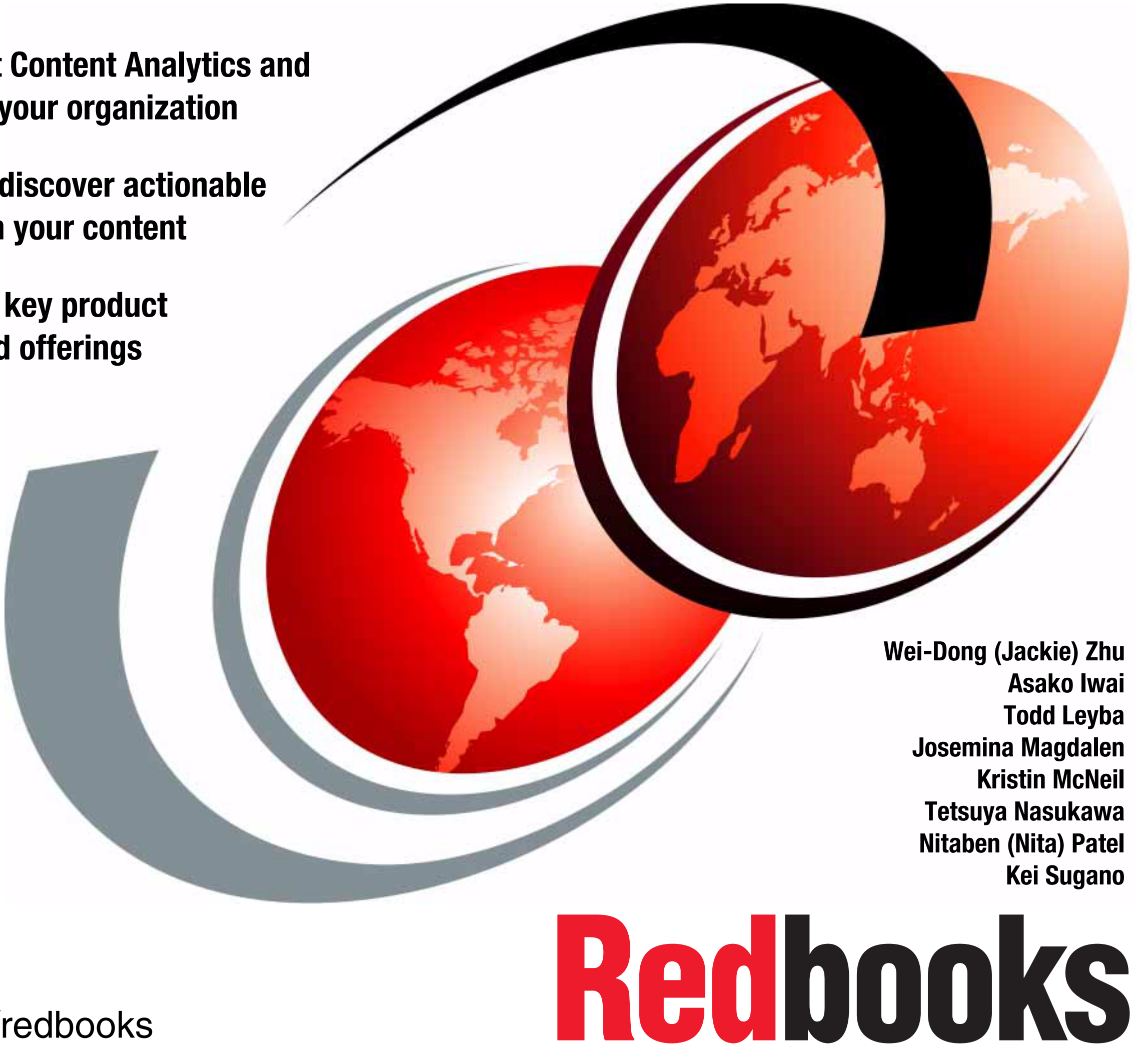
IBM Content Analytics Version 2.2

Discovering Actionable Insight from Your Content



- Learn about Content Analytics and its value to your organization
- See how to discover actionable insight from your content
- Understand key product features and offerings

Wei-Dong (Jackie) Zhu
Asako Iwai
Todd Leyba
Josemina Magdalen
Kristin McNeil
Tetsuya Nasukawa
Nitaben (Nita) Patel
Kei Sugano

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content

May 2011

Note: Before using this information and the product it supports, read the information in "Notices" on page xiii.

Second Edition (May 2011)

This edition applies to Version 2, Release 2, of IBM Content Analytics (program number 5724-Z21).

Contents

Notices xiii
Trademarks xiv

Preface xv
The team who wrote this book xvi
Now you can become a published author, too! xviii
Comments welcome xix
Stay connected to IBM Redbooks xix

Summary of changes xxi
May 2011, Second Edition xxi

Chapter 1. Overview of IBM Content Analytics 1
1.1 Business need and the Content Analytics solution 2
   1.1.1 Business need and problem statement 2
   1.1.2 The Content Analytics solution 2
1.2 History, changes, and what's new 3
   1.2.1 Product history 3
   1.2.2 Product changes 4
   1.2.3 What's new in Content Analytics Version 2.2 6
1.3 Important concepts and terminology 9
   1.3.1 Unstructured and structured content 9
   1.3.2 Text analytics 10
   1.3.3 Search, discovery, and data mining 10
   1.3.4 Collections 11
   1.3.5 Facets 12
   1.3.6 Frequency 13
   1.3.7 Correlation 13
   1.3.8 Deviation 15
1.4 Content Analytics architecture 15
   1.4.1 Main components 16
   1.4.2 Data flow 21
   1.4.3 Scalability 24
   1.4.4 Security 25

Chapter 2. Application design and preparation 27
2.1 Use-case scenarios 28
   2.1.1 Call center 28
   2.1.2 Insurance fraud 28

2.1.3 Quality assurance 29
2.1.4 Content assessment 29
2.2 Data considerations. 31
2.2.1 Content Analytics data model 31
2.2.2 Structured and unstructured sources 32
2.2.3 Multiple data sources 33
2.2.4 Date-sensitive data 34
2.2.5 Extracting information from textual data 35
2.2.6 The number of collections to use 36
2.3 Design guide for building a text analytics collection 36
2.3.1 Building a text analytics collection. 37
2.3.2 A walk through the building process 37
2.3.3 Planning for iteration 40
2.4 Programming interfaces 41
2.4.1 Search and Index API 41
2.4.2 REST API 43

Chapter 3. Understanding content analysis 45
3.1 Basic concepts of Content Analytics 46
3.1.1 Manual versus automated analysis. 46
3.1.2 Frequency versus deviation 49
3.1.3 Precision versus recall 51
3.2 Typical cycle of analysis with Content Analytics 53
3.2.1 Setting the objectives of the analysis 53
3.2.2 Gathering data 54
3.2.3 Analyzing data. 54
3.2.4 Taking action based on the analysis. 59
3.2.5 Validating the effect. 60
3.3 Successful use cases 60
3.3.1 Voice of customer 60
3.3.2 Analysis of other data 67
3.4 Summary 69

Chapter 4. Installing and configuring IBM Content Analytics. 71
4.1 Installing Content Analytics 72
4.1.1 Process overview 72
4.1.2 Confirming the system requirements and supported data sources. . 72
4.1.3 Determining the installation server type and procedure consideration73
4.1.4 Parameters used during the installation 74
4.1.5 Installing the agent server. 76
4.1.6 Installing Content Analytics on a single server 77
4.1.7 Running the First Steps program to verify the installation. 82
4.1.8 Starting the Text Analytics tutorial. 84

4.2 Administering Content Analytics 84
4.2.1 Starting the system 85
4.2.2 Accessing the administration console 85
4.2.3 Stopping the server 86
4.3 Configuring a text analytics collection 86
4.3.1 Designing a sample collection. 87
4.3.2 Creating a text analytics collection 90
4.3.3 Defining and configuring a crawler 123
4.3.4 Building an index in the text analytics collection 132
4.4 Verifying that the collection is available. 138
4.4.1 Starting the search server for the text analytics collection 139
4.4.2 Accessing the collection with the text miner application 139
4.5 Deploying the configuration. 139
4.5.1 Using the esadmin export and import commands. 139
4.5.2 Using the esbackup and esrestore commands. 140
4.5.3 Usage guidelines for these commands. 141

Chapter 5. Text miner application: Basic features 143
5.1 Overview of the text miner application 144
5.1.1 Accessing the text miner application. 144
5.1.2 Application window layout and functional overview 145
5.1.3 Selecting a collection for analysis 150
5.1.4 Changing the default behavior by using preferences 150
5.2 Search and discovery features 152
5.2.1 Limiting the scope of your analysis using facets. 153
5.2.2 Limiting the scope of your analysis using search operators 154
5.2.3 Limiting the scope of your analysis using dates 155
5.2.4 Query syntax. 156
5.2.5 Type ahead 157
5.2.6 Saved searches 161
5.2.7 Advanced search 162
5.3 Query tree 163
5.3.1 Accessing the query tree. 164
5.3.2 Understanding the query tree 165
5.3.3 Query tree examples. 166
5.3.4 Editing the query tree 177
5.4 Query builder. 180
5.4.1 Accessing the query builder 180
5.4.2 Features of the Query Builder window 182
5.4.3 Using the query builder 186
5.4.4 Preferred practice for using the query builder and the query tree. . 193
5.5 Rule-based categories with a query 193
5.5.1 Enabling the rule-based categories feature 194

5.5.2 Configuring rules for rule-based categories 194
5.5.3 Configuring rule-based categories 195
5.5.4 Adding the current query as a category rule 198
5.6 Common view features 203
5.7 Document flagging 205
5.7.1 Configuring document flags 205
5.7.2 Setting document flags 209
5.7.3 Viewing the document values of a flag facet. 213

Chapter 6. Text miner application: Views 217
6.1 Views. 218
6.2 Documents view 219
6.2.1 Understanding the Documents view 220
6.2.2 Viewing the document contents and facets. 221
6.2.3 When to use the Documents view. 222
6.3 Facets view 222
6.3.1 Understanding the Facets view. 225
6.3.2 Using the Facets view 225
6.4 Time Series view. 226
6.4.1 Features in the Time Series view 227
6.4.2 Understanding the Time Series view 230
6.4.3 Using the Time Series view. 231
6.5 Trends view. 231
6.5.1 Features in the Trends view 233
6.5.2 Sort criteria 235
6.5.3 Understanding the Trends view 236
6.5.4 When to use the Trends view 237
6.6 Deviations view 238
6.6.1 Features in the Deviations view 239
6.6.2 Understanding the Deviations view. 240
6.6.3 Using the Deviations view. 243
6.7 Facet Pairs view 244
6.7.1 Table view. 245
6.7.2 Grid view 245
6.7.3 Bird’s eye view 246
6.7.4 Understanding the Facet Pairs view with correlation values. 248
6.7.5 Using the Facet Pairs view 249
6.8 Connections view 250
6.8.1 Features in the Connections view. 252
6.8.2 Understanding the Connections view 257
6.8.3 When to use the Connections view. 261
6.9 Dashboard view. 262
6.9.1 Configuring the Dashboard 263

6.9.2 Viewing the Dashboard 274
6.9.3 Working with the Dashboard 275
6.9.4 Saving Dashboard charts as images 277

Chapter 7. Performing content analysis 279
7.1 Discovering actionable insight with the text miner application 280
7.1.1 The sample data 280
7.1.2 Insights without customization 283
7.1.3 Considerations about what you want to discover from the data . . . 286
7.2 Content analysis scenarios 287
7.2.1 Scenario 1: Using a custom dictionary to discover package-related calls 287
7.2.2 Scenario 2: Using custom text analysis rules to discover trouble-related calls. 291
7.2.3 Scenario 3: Discovering the cause of increasing calls 296
7.2.4 Conclusion 298
7.3 Configuring the Dictionary Lookup annotator 299
7.3.1 When to use the Dictionary Lookup annotator 299
7.3.2 Configuring custom user dictionaries 300
7.3.3 Migrating the Content Analyzer dictionaries 308
7.3.4 Validation and maintenance 308
7.4 Configuring the Pattern Matcher annotator 309
7.4.1 When to use the Pattern Matcher annotator 310
7.4.2 Configuring custom text analysis rules 310
7.4.3 Migrating the Content Analyzer rules 313
7.4.4 Designing the custom text analysis rules 313
7.4.5 Validation and maintenance 319
7.5 Preferred practices 320

Chapter 8. Discovering insight with terms of interest and document clustering. 321
8.1 The power of dictionary-driven analytics. 322
8.1.1 Multiple viewpoints for analyzing the same data. 322
8.2 Terms of interest 326
8.2.1 Basic algorithm for identifying terms of interest 330
8.2.2 Limitations in using automatic identification of terms of interest . . . 334
8.2.3 Preferred use of terms of interest identified automatically 335
8.2.4 Efficient and effective creation of dictionary 339
8.3 Document clustering 343
8.3.1 Setting up document cluster 343
8.3.2 Creating a cluster proposal 346
8.3.3 Refining the cluster results 348
8.3.4 Deploying clusters to a category 351

8.3.5 Working with the cluster results 353
8.3.6 Creating and deploying the clustering resource 354
8.3.7 Preferred practices 356

Chapter 9. Content analysis with IBM Classification Module 357
9.1 The Classification Module annotator. 358
9.1.1 When to use the Classification Module annotator. 358
9.1.2 The Classification Module technology. 359
9.2 Fine-tuning your analysis with the Classification Module annotator 361
9.2.1 Building your collection 361
9.2.2 Refining the analysis. 363
9.2.3 Using a conceptual search for advanced content discovery 364
9.3 Creating and deploying the Classification Module resource 365
9.3.1 Starting the Classification Module server 365
9.3.2 Creating and training the knowledge bases 366
9.3.3 Creating a decision plan 371
9.3.4 Deploying the knowledge base and decision plan 375
9.3.5 Configuring the Classification Module annotator. 378
9.4 Validation and maintenance of the Classification Module annotator. . . . 383
9.4.1 Using the Classification Module sample programs 384
9.4.2 Classification Module annotator validation techniques 384
9.5 Preferred practices 384

Chapter 10. Importing CSV files, exporting data, and performing deep inspection 387
10.1 Importing CSV files 388
10.2 Overview of exporting documents and data 395
10.2.1 Crawled documents 397
10.2.2 Analyzed documents. 398
10.2.3 Search result documents 398
10.2.4 Exported data manifest 399
10.3 Location and format of the exported data 400
10.3.1 Location of the exported data 400
10.3.2 Metadata format 402
10.3.3 Binary content format 403
10.3.4 Common Analysis Structure format 404
10.3.5 Extracted text format. 404
10.4 Common configuration of the export feature 405
10.4.1 Document URI pattern 405
10.4.2 Exporting XML attributes and preserving file extensions 405
10.4.3 Adding exported documents to the index 405
10.4.4 Exporting information about deleted documents. 406
10.4.5 Scheduling 406

10.5 Monitoring export requests 407
10.6 Enabling export and sample configurations 408
10.6.1 Exporting crawled documents to a file system for Content Collector 409
10.6.2 Exporting analyzed documents to a relational database. 415
10.6.3 Exporting search result documents to the file system for Classification Module 420
10.6.4 Exporting search result documents to CSV files 427
10.7 Deep inspection. 431
10.7.1 Location and format of the exported data 433
10.7.2 Common configuration 434
10.7.3 Enabling deep inspection 437
10.7.4 Generating deep inspection reports 438
10.7.5 Optional: Scheduling a deep inspection run 441
10.7.6 Monitoring the deep inspection requests 444
10.7.7 Validating the deep inspection reports generation 445
10.8 Creating and deploying a custom plug-in 446

Chapter 11. Configuring annotators 449
11.1 Document processing pipeline and the annotators 450
11.1.1 UIMA document processing pipeline. 450
11.1.2 Language Identification annotator. 450
11.1.3 Linguistic Analysis annotator. 451
11.1.4 Named Entity Recognition annotator 452
11.1.5 Dictionary Lookup and Pattern Matcher annotators 452
11.1.6 Classification Module annotator 452
11.2 Custom annotators 453
11.2.1 LanguageWare Resource Workbench 454
11.2.2 Creating custom annotators using the Apache UIMA SDK. 468
11.3 Validation. 476
11.3.1 Real-time NLP. 476
11.3.2 Advanced techniques 480
11.3.3 Summary of validation techniques 482

Chapter 12. IBM Content Assessment scenario. 483
12.1 Content Assessment offering 484
12.1.1 Concepts and terminology 485
12.2 Overview of Content Assessment 486
12.2.1 Content decommissioning scenario 487
12.2.2 Dynamically analyzing and collecting your content. 491
12.3 Content Assessment workflow 491
12.3.1 Decommissioning content. 492
12.3.2 Performing dynamic analysis 502

12.3.3 Preserving and using business data 507
12.4 Records management and email archiving 522
12.5 Preferred practices 523
12.6 Summary 524

Chapter 13. Integrating Cognos Business Intelligence 525
13.1 Initial setup 526
13.1.1 Running the esrepcog command 526
13.1.2 Configuring default application user roles 527
13.1.3 Configuring database connectivity 528
13.2 Generating Cognos BI reports. 533
13.3 Creating custom Cognos 8 BI reports 539
13.3.1 Configuring export options 539
13.3.2 Exporting search results 546
13.3.3 Loading the exported data model into Cognos 548

Chapter 14. Customizing and extending the text miner application . . . 557
14.1 Customizing the text miner application 558
14.1.1 Analytics Customizer. 558
14.1.2 Modifying the URI link in the Documents view 559
14.2 Reasons for extending the text miner application 560
14.3 Sample plug-ins for text miner views. 561
14.4 Customizing the sample text miner plug-in 562
14.4.1 Changing the view tab title 563
14.4.2 Customizing the plug-in template HTML file 563
14.4.3 Customizing the javascript widget. 564
14.4.4 Updating the style sheet for the plug-in. 567
14.5 Testing the customized plug-in 568

Chapter 15. Performance tips. 571
15.1 General performance guidelines 572
15.1.1 Factors that influence the performance of the system 572
15.1.2 Variables 574
15.2 Tuning the crawler component 576
15.2.1 Increasing active crawler threads 576
15.2.2 Setting the maximum heap size 578
15.3 Tuning the document processor 579
15.3.1 Setting the number of document processor threads 580
15.3.2 Increasing the maximum heap size. 581
15.4 Tuning the indexer. 581
15.4.1 Setting the number of indexer threads 582
15.4.2 Specifying the taxonomy cache type. 584
15.4.3 Increasing the maximum heap size. 585
15.4.4 Increasing the buffer size 587

15.4.5 Increasing the index commit interval 587
15.5 Enhancing the search performance 589
15.5.1 Increasing the search result cache entries 590
15.5.2 Enabling the optional facet index 591
15.5.3 Increasing the maximum heap size. 593
15.5.4 Setting the number of threads for rebuilding optional facet index . 594
15.6 Scalability 595
15.7 Monitoring the system 595
15.7.1 General guideline for monitoring the system. 596
15.7.2 Using the esadmin command utility 596
15.7.3 General guidelines for monitoring the operating system 601

Chapter 16. Hints and tips for troubleshooting 603
16.1 Overview of troubleshooting 604
16.2 General troubleshooting guidelines. 605
16.2.1 Understanding the problem. 606
16.2.2 Understanding the environment 607
16.3 Working with the logs in Content Analytics 607
16.3.1 The location of the logs. 608
16.3.2 Understanding the log contents 609
16.3.3 The first log to examine. 609
16.4 Installation and administration-related troubleshooting. 610
16.5 Text miner application-related troubleshooting 615
16.6 Data processing flow-related troubleshooting 618
16.7 Export-related troubleshooting 624
16.8 Classification Module server-related troubleshooting tips. 626
16.9 Reporting a problem to the IBM Software Support 627
16.10 Advanced troubleshooting topics 628
16.10.1 Changing the log level of the system and collection logs 628
16.10.2 Generating a javacore or a heapdump for a Java session 629

Appendix A. Security in IBM Content Analytics. 633
The security concept in Content Analytics 634
Authentication 634
Authorization 634
Administrative access control 635
Collection-level access control 636
Document-level access control 636
Data encryption 638
Enabling login security in the embedded application server (Jetty) 639
Configuring application user roles 642
Configuring application user roles for the system 642
Configuring application user roles for specific user or group 644

Limiting user access to the text analytics collection 647
Task overview 647
Scenario overview 647
Mapping an application ID with collections and the user 648
Verifying security privileges. 650
Managing an application ID. 651

Related publications 653
IBM Redbooks 653
Other publications 653
Online resources 653
Help from IBM 654

Index 655

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the web at `http://www.ibm.com/legal/copytrade.shtml`

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®
alphaWorks®
Cognos®
DB2®
Domino®
DS4000®
FileNet®
GPFS™
IBM®
InfoSphere™
LanguageWare®
Lotus Notes®
Lotus®
MVS™
Notes®
OmniFind®
POWER®
Redbooks®
Redbooks (logo) ®
WebSphere®

The following terms are trademarks of other companies:

Adobe, the Adobe logo, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Java, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Preface

With IBM® Content Analytics Version 2.2, you can unlock the value of unstructured content and gain new business insight. IBM Content Analytics Version 2.2 provides a robust interface for exploratory analytics of unstructured content. It empowers a new class of analytical applications that use this content.

Through content analysis, IBM Content Analytics provides enterprises with tools to better identify new revenue opportunities, improve customer satisfaction, and provide early problem detection. Content Analytics offers the following key capabilities:

- *Discover* new relationships in enterprise content.
- *Refine* content to provide business context using semantic and faceted search capabilities.
- *Deliver* new insights to business users, applications, or business processes and convert that insight into active, focused decision making.

To help you achieve the most from your unstructured content (also referred as *textual content*), this IBM Redbooks® publication provides in-depth information about Content Analytics. This book examines the power and capabilities of Content Analytics, explores how it works, and explains how to design, prepare, install, configure, and use it to discover actionable business insights.

This book explains how to use the automatic text classification capability, from the IBM Classification Module, with Content Analytics. It also explains how to use the LanguageWare® Resource Workbench to create custom annotators. In addition, it explains how to work with the IBM Content Assessment offering to timely decommission obsolete and unnecessary content while preserving and exploiting content that has business value.

The target audience of this book is decision makers, business users, and IT architects and specialists who want to understand and use their enterprise content to improve and enhance their business operations. It is also intended as a technical guide for use in conjunction with the online information center for configuring and performing content analysis with Content Analytics.

IBM Cognos reference: IBM Content Analytics was previously and briefly known as *IBM Cognos Content Analytics*. In this book, you will notice the previous name in some of the application windows and in the references to the information center for IBM Content Analytics.

The team who wrote this book

This book was produced by a team of specialists from around the world working at the International Technical Support Organization (ITSO).

Wei-Dong (Jackie) Zhu is an Enterprise Content Management Project Leader with ITSO. Jackie joined IBM in 1996 and has more than 10 years of software development experience in accounting, image workflow processing, and digital media distribution. She is a Certified Solution Designer for IBM Content Manager and has managed and lead the production of many Enterprise Content Management Redbooks publications. Jackie holds a Master of Science degree in Computer Science from the University of the Southern California.

Asako Iwai is a member of the IBM Product Support team for content discovery products including enterprise search. She joined IBM in 2000 and worked as a member of the IBM DB2® Linux®, UNIX®, and Microsoft Windows® Product Support team for six years supporting customers in Japan. She is a DB2 Linux, UNIX, and Windows Certified Advanced Database Administrator. Asako moved to content discovery worldwide customer support in 2007 and has three years of experience providing support for the IBM OmniFind® Enterprise Edition product. Asako holds a Master of Science degree in Applied Physics from Waseda University in Japan.

Todd Leyba is an architect of search and text analytic products in the content discovery organization of IBM Information Management division. Todd joined IBM in 1986 and, since then, has worked on a variety of search and text analytic-related projects. Such projects include IBM Content Analytics, IBM Classification Module, IBM OmniFind Enterprise Search, IBM Infomarket Service, Mantis (IBM's first crawling and indexing technology), and IBM Lotus® Extended Search. He holds a bachelor degree from the University of Maryland and a Master of Science degree in Computer Science from Johns Hopkins University.

Josemina Magdalen is a Senior Team Leader and Architect for the IBM Israel Software Group (ILSL). She has a background in Natural Language Processing (including text classification and search) and in text mining and filtering technologies. Josemina joined IBM in 2005 and has worked in the Content Discovery Engineering Group doing software development projects in text categorization, filtering and search, as well as text analytics. Prior to joining IBM, Josemina has worked in Natural Languages Processing research and development (machine translation, text classification and search, data mining), for over ten years. Josemina co-authored the Redbooks publication *IBM Classification Module: Make It Work for You*, SG24-7707.

Kristin McNeil is an IT Specialist with the Software Group in the US. She has more than 10 years of software development experience. Her areas of expertise include quality assurance for IBM Content Analytics and OmniFind Enterprise Edition. Kristin holds a Masters of Business Administration in Management Information Sciences degree from State University of New York at Albany.

Tetsuya Nasukawa is a research staff member of IBM Research in Tokyo, Japan. He joined IBM in 1989, and he has been leading text mining projects since 1997. Tetsuya is the primary inventor of the Text Analysis and Knowledge Mining (TAKMI) system that has been integrated into Content Analytics. He has 25 years of experience in the natural language processing field. His areas of expertise include text mining, machine translation, sentiment analysis, and conversation mining. Tetsuya has authored and co-authored more than 50 academic papers and received 8 academic awards. He has also written several books in Japanese and wrote the text mining section of Encyclopedia of Natural Language Processing (Japanese). Tetsuya holds a PhD. degree in engineering from Waseda University, Japan.

Nitaben (Nita) Patel is a member of the Quality Assurance (QA) team for IBM Content Analytics. She joined IBM in 2004 as part of Venetica acquisition. She has more than 7 years of testing experience in technologies related to enterprise content management, content integration, and search and text analytics. She is also part of QA Architect team that focuses on areas related to performance test, integration test, and test automation. Nita holds a Master of Science degree in Information Technology from the University of the North Carolina.

Kei Sugano is a software engineer at the IBM Yamato Software Development Laboratory in Japan. He joined IBM in 2004 and has been working on the development of linguistic components and resources as a member of the IBM LanguageWare team. He has more than 5 years of experience in developing Unstructured Information Management Architecture (UIMA)-compliant analysis engines and supporting UIMA-based products including IBM OmniFind Enterprise Edition, IBM Content Analyzer, IBM InfoSphere™ eDiscovery Analyzer, and IBM Content Analytics. Kei holds a Master of Science degree in Mathematics from the Hokkaido University, Japan.

A special thank you to Teruo Koyanagi for allowing us to base our performance chapter on his performance white paper.

Special thanks to the following people for making this book production possible:

Seiji Hamada
Takeshi Inagaki
An Nguyen

We also thank the following people for their contributions to this project. We could not produce the book without their help and assistance:

Noritaka Adachi
Shunsuke Ishikawa
Hiroaki Kikuch
Masaki Komedani
Teruo Koyanagi
Hiroshi Kurokawa
Qiu Hui (Ivy) Ling
Kristin McNeil
Yutaka Moriya
Takuma Murakami
Keisuke Nitta
Vinod Reddy
Gang (James) Zhan
Yamato Software Development Laboratory (YSL), SWG, IBM Japan

Patrick Fink
Rashmi Vittal
Peggy Zagelow
IBM Silicon Valley Lab, San Jose, California

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

`ibm.com/redbooks/residencies.html`

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-7877-01
for IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content
as created or updated on May 10, 2011.

May 2011, Second Edition

This second edition is updated to include the most important new features in IBM Content Analytics Version 2.2. It is not a comprehensive update to the previous release. This section outlines the new topics covered in this edition and changes to the previous edition.

New topics covered

The following new topics have been added to this book for IBM Content Analytics Version 2.2:

- Text miner application views (Chapter 6, “Text miner application: Views” on page 217):
  - Connection View. The new graphical view shows the relation of the two selected facets to help you visualize the connection between them.
  - Dashboard View. With this new view, you can select multiple views in the text miner application and see them in a single dashboard view.
- New text miner application user interface changes (other than the views):
  - Type ahead (Chapter 5, “Text miner application: Basic features” on page 143)
  - Facet ranges (Chapter 4, “Installing and configuring IBM Content Analytics” on page 71, and Chapter 5, “Text miner application: Basic features” on page 143)
  - Document flagging (Chapter 5, “Text miner application: Basic features” on page 143)

- New discovering insight features (Chapter 8, “Discovering insight with terms of interest and document clustering” on page 321):
  - Terms of interest. With this feature, the system can automatically provide possible terms of interest for further investigation.
  - Document clustering. With this feature, the system can provide cluster suggestions for documents that can be used for document classification for easier search and analysis.
- Comma-separated values (CSV) file import (Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387)

  The system can import CSV files so that content can be quickly added to a collection without setting up a crawler or accessing a content repository.
- Cognos BI Integration (Chapter 13, “Integrating Cognos Business Intelligence” on page 525)

  Information has been added that explains the seamless integration between IBM Content Analytics and Cognos 8 BI.
- Extending the text miner application (Chapter 14, “Customizing and extending the text miner application” on page 557)

  An example has been added that explains how to add one or more of your own text analytic views for customization to meet specific data and visualization needs.

Other changes to the previous version

This edition includes revisions to the existing chapters with the latest window captures and key information related to Version 2.2. The chapters were reorganized to better accommodate the new features and functions of the product.

Some specific changes include the following areas:

- Installation and configuration
- The basics of the text miner applications (Query Tree and Query Builder)
- Content analysis with classification module
- Export data
- LanguageWare Resource Workbench integration
- Content Assessment scenario
- Performance tuning
- Troubleshooting hints and tips
- Security

New information

The following existing chapters have been enhanced with the following new information:

- Chapter 1, “Overview of IBM Content Analytics” on page 1, now includes a consolidated list of the new features and functions for IBM Content Analytics Version 2.2. Also, the architecture topic now includes information about improved scalability features.
- Chapter 2, “Application design and preparation” on page 27, includes minor editorial changes to account for the new features and functions listed in Chapter 1, “Overview of IBM Content Analytics” on page 1. This chapter also introduces the provision of a new REST-based API.
- Chapter 4, “Installing and configuring IBM Content Analytics” on page 71, now includes the configuration rationale and steps for the following new capabilities:
  - Importing CSV files
  - Terms of interest
  - Query type ahead
  - Input field filtering
  - Configuration of category tree facets
- Chapter 5, “Text miner application: Basic features” on page 143, has been expanded with the addition of a new Chapter 6, “Text miner application: Views” on page 217, which now contains the following new topics:
  - Connections view
  - Query builder
  - Concept based searching
  - Near duplicate document detection
  - New features added to the Query Tree
- Chapter 7, “Performing content analysis” on page 279 (previously Chapter 6 in the first edition), includes a section that briefly describes the designing of the custom text analysis rules.
- Chapter 8, “Discovering insight with terms of interest and document clustering” on page 321, is new and includes both term-of-interest and document clustering features.
- Chapter 9, “Content analysis with IBM Classification Module” on page 357 (previously Chapter 7 in the first edition), explains how the IBM Classification Module is used to support concept based searching in IBM Content Analytics.
- Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387 (previously Chapter 8 in the first edition), now includes a section on importing and exporting data to a CSV file.

- Chapter 11, “Configuring annotators” on page 449 (previously Chapter 9 in the first edition), has been updated to reflect the tighter integration with the LanguageWare Resource Workbench that is used to build and test custom text analytics. It also explains how to use the new REST API with LanguageWare.
- Chapter 12, “IBM Content Assessment scenario” on page 483 (previously Chapter 10 in the first edition), has been extended. It now explains how the new document flagging, document clustering, and terms-of-interest features can be used in the content assessment scenario.
- Chapter 13, “Integrating Cognos Business Intelligence” on page 525, is new and explains the seamless integration between IBM Content Analytics and Cognos 8 BI.
- In Chapter 14, “Customizing and extending the text miner application” on page 557, the customization information from the first edition was consolidated with the new extending example.
- Chapter 15, “Performance tips” on page 571 (previously Chapter 11 in the first edition), explains a few new techniques that you can use to increase the overall performance of your IBM Content Analytics system.
- Chapter 16, “Hints and tips for troubleshooting” on page 603 (previously Chapter 12 in the first edition), has been updated with the latest information. Some obsoleted information was removed.
- Appendix A, “Security in IBM Content Analytics” on page 633, now includes information about how to add user application roles to control access to specific text miner application functions.

1

Overview of IBM Content Analytics

With major advances in linguistic analysis of the written word, combined with the increased computational power of today's hardware, comes the field of *text analytics*. Text analytics enables businesses to gain insight and understanding from their textual data (also often referred as *unstructured content*). IBM Content Analytics Version 2.2 is the product in this emerging market that employs text analytics so that businesses can make the best use of all of their content.

This chapter includes the following sections:

- Business need and the Content Analytics solution
- History, changes, and what's new
- Important concepts and terminology
- Content Analytics architecture

1.1 Business need and the Content Analytics solution

This section introduces you to the Content Analytics product. First it explains the business need and problem statement. Then it provides the Content Analytics solution to the problem followed by a brief history of the product. Lastly this section describes what is new in release 2.2 of Content Analytics.

1.1.1 Business need and problem statement

A large percentage (estimated at 80% or more) of the information for a company is maintained as textual data, which includes valuable assets such as emails, free-form fields on applications, wikis, and text messages. Because this content lacks structure, it is difficult to analyze it with automation. To understand and analyze the content, generally a human must read and understand what is being communicated. As a consequence, human involvement can be an expensive and time consuming component of your overall business process.

Applying content analytics to your textual data through the help of software can result in many benefits. For example, analytics applied to online customer postings can help target and deliver new branding campaigns, increasing sales and customer loyalty. Analytics applied to text in insurance claim files can help detect fraud faster, reducing costs for your clients and optimizing the claims handling process. Analytics applied to customer comments and feedback, also known as *voice of customer* (VOC), provides insight to drive customer-oriented decision making, boosting loyalty and creating new opportunities.

1.1.2 The Content Analytics solution

The Content Analytics product is designed to process your textual data in ways that help you to search, discover, and perform the same analytics on textual data that is currently performed on structured data. With Content Analytics, you can now use your text in ways that were only previously attainable from your structured data.

Content Analytics delivers new business understanding and visibility from the content and context of textual information. For example, you can identify patterns, view trends over time, and reveal unusual correlations or anomalies. You can explain why events are occurring and find new opportunities by aggregating the voices of customers, suppliers, and the market. You can track and drive improvement in non-quantitative business metrics through content dashboards, reports, and scorecards. In addition, Content Analytics helps reduce costs by exposing irrelevant or obsolete content for deletion (also known as *content decommissioning*). For more information about content decommissioning, see 12.2.1, “Content decommissioning scenario” on page 487.

With Content Analytics, you can also analyze your textual data when it is not practical to analyze it manually. For example, if you conduct a survey with 1,000,000 people on what they do over the weekend, Content Analytics helps you to analyze all 1,000,000 survey forms.

With Content Analytics, you can define many *facets* (or aspects) of your data, with each facet potentially leading to valuable insights for various users. For example, you might define a weekend destination facet that consists of major places where people travel over the weekend. You might also define an activity facet that consists of typical activities people do during their weekend travel. With such facets, a tourist industry analyst can analyze which types of people (based on their age, profession, gender, and other aspects) tend to travel to which specific locations. You can further identify the types of activities they engage in over the weekend.

In another example, you might define a shopping place facet that consists of major places for shopping and a purchase facet that consists of items being purchased by people. With these facets, retailers can analyze the type of people that tend to buy particular items at a given location.

Content Analytics is tool for reporting statistics and for obtaining *actionable insights*. Actionable insights is a key concept that refers to insight into data that leads to action. Content Analytics provides far more value than just being a tool to reduce the workload of manual analysis. Content Analytics brings the power of business intelligence to all of your enterprise information, not just your structured information. The result helps you achieve the most value from all your data, regardless of its structure.

1.2 History, changes, and what’s new

Content Analytics has gone through several revisions and changes. This section provides a brief history of the product, summarizes the changes in the product, and explains what is new in Version 2.2.

1.2.1 Product history

In 1997, IBM started a text mining project in IBM Research – Tokyo in Japan. It combined natural language processing (NLP) technology that was inherited from machine translation and digital library projects.

By 1998, the Text Analysis and Knowledge Mining (TAKMI) system was developed from the text mining project. TAKMI was used to analyze 500,000 customer contact records at a PC help center in Japan. A corresponding help center in the United States started using the English version of the TAKMI system in 1999. The help center reported a significant cost reduction based on identification of product failures in their early stages with the TAKMI system.

From 2000, large enterprises started using the TAKMI system both in Japan and the US. To protect their competitive advantage, most of these companies kept their use of the TAKMI system confidential. As a direct result of using the TAKMI system, the PC help center in Japan achieved number one in problem solving ratio of web support among PC companies operated in Japan in 2003, as ranked by *Nikkei Personal Computing*, the premiere PC magazine in Japan.

The use of the TAKMI system has expanded greatly since 2003, primarily because of On Demand Innovation Services (ODIS), in which IBM Research directly supports client use of text mining by sending researchers as consultants.

In 2007, IBM Software Group released the TAKMI system as a Price Requested Quote (PRQ) service offering named *IBM OmniFind Analytics Edition*. OmniFind Analytics Edition received Spring 2008 SSPA Recognized Innovator Awards from the Service and Support Professionals Association (SSPA).

In 2008, IBM renamed OmniFind Analytics Edition to *IBM Content Analyzer*. In 2009, IBM developed Content Analytics by integrating technology from Content Analyzer.

In October 2009, version 2.1 (the first release) of Content Analytics debuted. Then in October 2010, version 2.2 of Content Analytics was released.

> **Cognos reference:** Content Analytics was also briefly known as *IBM Cognos Content Analytics*. In this book, you will notice the previous name in some of the application windows and in the references to the information center for Content Analytics.

1.2.2 Product changes

The integration of the Content Analyzer product into the Content Analytics product resulted in changes in the terminology and concepts used. This section reviews these changes for those users who are familiar with OmniFind Analytics Edition or Content Analyzer. Table 1-1 on page 5 shows the change in terminology between the two products.

Table 1-1 Terminology change from Content Analyzer to Content Analytics

Content Analyzer	Content Analytics
Database	Collection
Category	Facet

Major enhancements

Content Analytics provides the following enhancements since the earlier product:

- All text analytics views are enhanced in both function and appearance.
- A new rich query syntax helps narrow down document sets.
- Supported languages are increased from 2 to 11.
- Documents can be in different languages within one particular collection.
- The GUI is easier to customize.

Notable differences

Table 1-2 lists the notable differences between Content Analytics and Content Assessment.

Table 1-2 Notable differences in Content Analytics

Item	Description
Process model	Content Analytics consists of several sessions. Each session serves a particular function and communicates with each other. The foundation is provided by the common communication layer (CCL) that controls the sessions.
Integrated application server	Content Analytics bundles its own application server (Jetty web server) to offer a web interface. IBM WebSphere® Application Server does not need to be separately. However, the text miner application can be deployed in WebSphere Application Server if necessary.
Administration GUI	Content Analytics provides a GUI application to administer the system.
Character normalization	Content Analyzer applies language-dependent character normalization. However, Content Analytics applies common normalization, which is Unicode NFKC normalization with few custom additional rules. The result of normalization between Content Assessment and Content Analytics is different.
Data input	Content Analytics provides various crawlers to collect data from data sources, for example, relational database (RDB), web, and file system. By using the crawler plug-in, you can integrate custom data cleansing logic within the crawling.

Item	Description
Moving data	Content Analytics provides import and export capability to copy collection configuration. Content Analytics also provides a backup and restore capability that duplicates both configuration and data.
Scalability	Content Analytics supports multinode configuration for high scalability.
Incremental update	Content Analytics can process additional input documents or rebuild data without stopping the Text Analytics services.

1.2.3 What's new in Content Analytics Version 2.2

Content Analytics Version 2.2 has greatly extended the features and functions of Content Analytics. These new capabilities benefit business analysts and assist the text analytics and systems administrators of Content Analytics. The new features for each are briefly described in the following sections.

New features for the business analyst

Business analysts can take advantage of the following new features:

- Several new text miner views are provided:
  - The *Connections view* is similar to the facets pairs view. This new view compares two selected facets and shows the correlation between their values in the form of a connected graph. The nodes in the graph represent the values of the facets and the arcs that connect the nodes, indicating the degree of correlation between them.
  - The *Dashboard view* which, as the name implies, allows you to configure and display multiple text miner views into a single dashboard view. The dashboard view provides for a convenient way to obtain a single holistic view of your data.
  - A *customizable view* is now available with its content defined by you. With Content Analytics, you can now plug in your own view, which is displayed as a new tab in the text miner application. With this new feature, you can integrate other text analytic tools that you use on a regular basis into the text miner for your convenience.
- Content Analytics Version 2.2 also offers its greatly improved integration with the IBM Cognos BI Reporting module. Nearly all of the text miner views now provide a Cognos Report generation button, which automatically generates a Cognos report depicting the current state of your data as it is being analyzed in the text miner. You can view the Cognos report from within the text miner or

from the Cognos report module interface. Conversely, when viewing the report from Cognos, a link can take you back to the text miner restored to the state when the report was generated. This seamless integration with Cognos is a tremendous improvement from the integration offered in Version 2.1.

- Content Analytics Version 2.2 now supports document flagging, a feature with which you can group discrete search results and assign them to a common flag. You can export similarly flagged documents later or select them for further investigation.
- The new near duplicate document detection is an optional feature. When this feature is activated, you can see and group documents that Content Analytics has determined to be near duplicates of each other. The elimination of duplicates can greatly improve the accuracy of the calculations made by the system and your subsequent analysis.
- Content Analytics can automatically identify main topics of discussion in the text by measuring the statistical frequency and collocation of words among each other. Words that tend to appear frequently together more often than others are assigned to a cluster and can then be viewed by using the cluster facet in the text miner application. The clustering technology is provided by an embedded version of the IBM Classification Module, alleviating the need to separately purchase the IBM Classification Module product.
- You can now build your own facet category tree by using queries defined by you. Documents that match a category query during indexing are assigned to its corresponding facet category. In addition to specifying a query expression, the administrator can alternatively specify a Uniform Resource Identifier (URI) pattern.
- You can now automatically generate *terms of interest* that are encountered in your text. You can use these terms of interest later in the formation of custom dictionaries. The terms of interest are selected based on the natural language parsing of the text and the statistical usage of non-domain specific words encountered and their relationship to nouns and verbs. For more information about terms of interest, see Chapter 8, “Discovering insight with terms of interest and document clustering” on page 321.
- A new *query builder* and enhanced *query tree* are also now available in Version 2.2. With these features, you can easily express and build complex queries without having to know the specific syntax of the Content Analytics query language.
- *Type ahead* is a new feature that suggests a search query with the estimated numbers of results based on what you have typed so far into the search box. The suggested queries can be previous queries entered by you and others using Content Analytics, or they can be from terms in the index that start with the letters you have typed so far in the query box.

- Query Suggestion has been enhanced to support more languages and noun phrases. Query suggestions are shown after you submit your original query and is different from query type ahead.
- In Version 2.2, you can now use any number of date fields and switch between them in the text miner. This is an improvement over Version 2.1, where you can only use one date field in your data.
- With the new comma-separated value (CSV) format, you have a new option for saving search results. By using the CSV format, you can more easily share your result data with others by easily ingesting the data into common tools such as spreadsheets or relational databases.

New features for the text analytics administrator

Text analytics administrators can take advantage of the following new features in Content Analytics Version 2.2:

- You can now import CSV files for indexing and easier custom dictionary development.
- With enhanced security, you can control access to specific functions in the text miner application. You can map individual users, groups, or both to many different discrete functions.
- In addition to Content Analytics Java™ API Search and Index API (SIAPI), developers can now choose to use a Representation State Transfer (REST) API that is SearchRest 2.0 compliant. They can use this new API to develop both search and administrative applications.
- A new input field filter helps in the data cleansing effort. If a field matches a condition, for example equals a certain value or is null, you can map it to another field or replace it with another value., such as to map it to another field or replace it with another value.
- You can now configure facets as ranges for date and numeric types of facet values.
- Because of tighter integration with IBM LanguageWare Resource Workbench, you can perform highly customized annotator development and testing between the workbench and the Content Analytics product.

New features for the systems administrator

System administrators can take advantage of the following new features in Content Analytics Version 2.2:

- More flexible server node assignment supports the switching between server roles (for example, data processing and search runtime nodes)
- High Availability Configuration for Microsoft Windows and AIX® platforms support 24x7 operation. An identically configured Content Analytics master

server is placed on hot standby and is automatically activated when High Availability detects a hardware failure.

1.3 Important concepts and terminology

To understand the Content Analytics product, you must first understand the key concepts and terminology that are associated with the product and technology. This section explains the following concepts and terminology:

- Unstructured and structured content
- Text analytics
- Search, discovery, and data mining
- Collections
- Facets
- Frequency
- Correlation
- Deviation

You *must* read this section before proceeding to the rest of the book.

> **Fast path to content analysis:** If you are familiar with the concept and mechanical operation of Content Analytics and you are interested in content analysis immediately, use the following fast path:
>
> 1. Locate a working Content Analytics system, which can be a VMware image that has Content Analytics installed.
> 2. Configure the sample collection from the First Steps tutorial.
> 3. Read Chapter 3, “Understanding content analysis” on page 45, in its entirety.
> 4. Read and follow the instructions in Chapter 7, “Performing content analysis” on page 279.

1.3.1 Unstructured and structured content

Unstructured content is information that is generally recorded in a natural language as free text. The text contains all of the complexities and ambiguities of the language that is being used. It is easily understood by a human reader but difficult to process by a computer program. In contrast, *structured information* is data that has unambiguous values and is easily processed by a computer program.

For example, in a typical email, the from, to, and date fields contain structured data with implied rules about what they mean and how they are to be processed. The from and to fields are email addresses. The date field has a data value and conforms to one of a limited set of date formats. Conversely, the body of the email (a field itself) has no implied structure, only to the extent that the text conforms to the grammatical rules of a particular natural language. Even then, the variance in each user's style of writing cannot guarantee that all grammatical rules are followed precisely.

In this book, unstructured content is referred to as *textual data*.

1.3.2 Text analytics

Text analytics is a general term that refers to the automated techniques of converting textual data into structured data. A program that reads text and extracts person names is considered a text analytic. A program that classifies the content into one or more categories based on the text that was read is also a text analytic. After the information in the text is converted to a structured form, the data can then be processed by conventional business intelligence and data processing tools that are available.

1.3.3 Search, discovery, and data mining

Search and discovery are often times mistaken as having the same meaning or at least as being interrelated in some way. Actually, search and discovery are different concepts. One way to contrast the two is to classify them by what you know and do not know. You search for what you know and discover what you do not know.

When you *search*, you already have a target in mind, such as a document, product, or piece of information. Your task is to formulate your query in a way that improves the chances for an exact or partial match of the target document. *Keywords* in your query tend to be more descriptive to qualify exactly what you are looking for. For example, the query “replacement air filter for a car model <x> year <y>” leaves little room for ambiguity.

Keywords: As the term implies, keywords are usually words and phrases that are extracted from textual content. However, they can also be obtained from structured fields such as date or numeric fields.

Discovery is exploratory in nature and is generally goal driven. A search engine becomes a discovery engine when the query is used as a starting point from which to learn more about a particular topic.

Data mining is the process of identifying patterns in your data that might be used to answer a business problem, question, or concern. Data mining is a natural part of discovery. Many techniques can be used in the data mining process with the patterns revealed in many different ways.

Content Analytics embodies each of the concepts described previously. The product is primarily a data mining tool for textual content. It combines the results with those of your structured data. By using the text miner application, you can explore and mine your data regardless of where it comes from and whether it is structured or unstructured. Search is also integrated into the product, so that you can limit your analysis to only those documents that match your search query.

1.3.4 Collections

A single Content Analytics *collection* represents the entire group of documents that are available to an application for search and analysis. An application can access multiple collections. The entire group of documents within a collection is sometimes referred to as a *document corpus*.

You can set up your collection as a *search collection* or a text analytics collection. A search collection is set up for use in a case of search applications only. A *text analytics collection* is set up for use when discovery and data mining are required.

> **Text analytics collection:** The remainder of this book focuses solely on text analytics collections. See the *IBM Cognos Content Analytics, Version 2.2.0 Administration Guide*, SC19-2875, for more information about search collections.

Content Analytics supports the creation of multiple collections. Each collection has its own set of configuration files and processes, such as crawlers, document processors, an indexer, and search run times. A collection is empty until content is added to it through the definition and scheduling of one or more crawlers and subsequent parsing and indexing of the crawled content. Documents can also be pushed by using the SIAPI.

Additional configuration options are available when defining crawlers and configuring the parser, indexer, and search components. For more information about these components, see 1.4, “Content Analytics architecture” on page 15.

1.3.5 Facets

Facets represent the different aspects or dimensions of your document corpus. They are a crucial mechanism for navigating and analyzing your content with the text miner application.

Facets can be populated with values that are obtained directly from the structured data fields in your collection. For example, you might have a set of traffic accident reports. Each report records the time of day that the accident occurred, the road condition (dry, wet, or icy), the number of vehicles involved, and so on. Each report also contains a free-form text field where a detailed description of the accident is recorded.

You want to investigate whether the road condition is a factor in causing accidents. Therefore, you create a road condition facet in Content Analytics. You populate the facet with the values from the road condition field of the traffic accident report.

Facets can also be populated with information from your text. For example, you want to know what type of cars were involved in the accidents. Therefore, you create a car model facet. The traffic accident report does not contain a structured field for car model. However, this information might exist in the detailed description field. To identify and extract the car model value for the facet, you employ text analytics to the description field and obtain the value from it for further analysis.

A set of default facets is automatically defined and populated by Content Analytics. The default facets represent the parts of speech (such as nouns and verbs) discovered in your text. Phrase constituent is another default facet. These facets provide important dimensions to your data and are a useful tool in your analysis. Chapter 3, “Understanding content analysis” on page 45, shows examples of how to use these facets.

When used alone or combined with a search query, facets provide a powerful way to navigate and filter your corpus of documents so that you focus on only those documents that are relevant to your analysis. In the previous example of traffic accidents, consider that you only want to focus on accidents that occurred when the roads were wet. You can easily obtain this information by selecting the wet value in the road condition facet and adding it as a constraint to your query.

Let us further assume that, for this smaller set of accident reports, you also want to focus on just those accidents that involved a brake failure. A structured field that identifies brake failure does not exist in the accident report. However, you can still narrow your document set by adding a search query term, such as “brake failure,” to dynamically filter your documents.

Upon defining your facets and building a text analytics collection, all of your facets are displayed in the text miner application. Where the data came from is irrelevant after it is examined by the text miner application. Most important is that you now have a way to navigate through your documents by the various dimensions represented by the facets.

You can now pose queries regarding the following information:

- Show me the distribution and frequency of accidents across the different types of road conditions (using the road conditions facet).
- Show me the frequency of accidents across the different models of cars involved (using the model car facet).

More importantly you can compare facets to each other to determine if they are correlated. See 1.3.7, "Correlation" on page 13, for more information about correlations.

1.3.6 Frequency

Frequency counts in Content Analytics represent the total number of documents that contribute to a particular keyword. The frequency can change as your query constraints changes.

For example, by selecting the road condition facet described in 1.3.5, "Facets" on page 12, the text miner application shows the total number of documents (frequency) for each keyword: dry, wet, and icy. You can add terms, such as "brake failure," to the search query to further constrain the document set. In this case, the frequency counts are automatically updated to reflect only those document totals within the number of accidents reports that contain the words "brake failure" in their description field.

Frequency counts can be useful in identifying trends in your data. The text miner application, for example, shows whether the number of car accident reports are increasing or decreasing over time.

1.3.7 Correlation

Although frequency counts are useful, relying on them alone can sometimes be misleading. Because a particular keyword has a high frequency count does not mean that it is relevant to your analysis. Content Analytics also calculates a correlation statistic for facets.

Correlation indicates how two facets are correlated to each other. They are used to better gauge the relevance of a particular keyword as it compares to other data in your document corpus.

Let us consider how correlation values can be used in the traffic accident analysis. In this case, you want to focus on traffic accidents that occurred when the roads were wet. You have the car model facet and the road condition facet. You set the road condition facet to the “wet” value. Content Analytics automatically updates the correlation values of all facets as they compare to wet road conditions. The higher the correlation value is, the more correlated they are to each other. In this example, the higher the correlation value is, the more relevant the traffic accident is linked to the wet road condition.

Figure 1-1 shows all the traffic accidents reported for car model X and car model Y. It also shows the number of car accidents when the road condition is wet for both car models. In this diagram, the letters X and Y represent different models of cars involved in accidents. Each instance of a letter is a unique accident report for the car model. Thus you have the following statistics:

- A total of 20 traffic accidents involved car model X, 10 of which occurred when the roads were wet.
- A total of 9 traffic accidents involved car model Y, 7 of which occurred when the roads were wet.

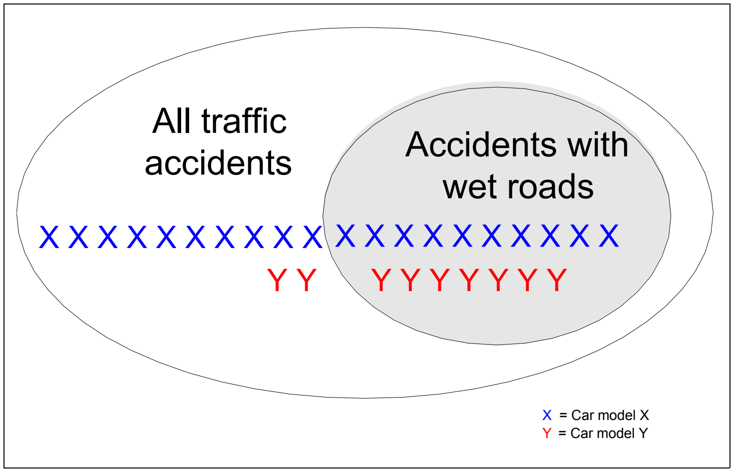


Figure 1-1 Car models correlated with a wet road condition

At first glance, based on the frequency of accidents alone, you might think that car model X has a problem when the roads are wet because traffic accidents occurred 10 times when the roads were wet for car model X. However, upon further examination, you notice that car model X has the same number of accidents regardless of whether the roads were wet. This means that a wet road condition might not be statistically relevant for the traffic accidents for car model X. The high number of accident reports for car model X (when the road was wet) is probably because car model X is a popular selling car and more of them are on the road.

Using the correlation value calculated by Content Analytics, car model Y has a bigger problem when the roads are wet. People who drive car model Y tend to have higher traffic accidents when roads are wet (7 related to a wet condition in a total of 9 traffic accidents). The correlation of wet road conditions and car accidents for car model Y is much higher than for car model X.

A high correlation value is important and is worth further investigation. In the example, the high correlation value indicates that a problem might exist with car model Y when the roads are wet. The problem can contribute to brake failure or an electrical problem caused by wet road conditions. Therefore, further investigation is recommended.

1.3.8 Deviation

In addition to frequency and correlation statistics, Content Analytics can identify trends and patterns that occur over time. The time is based on one or more date fields that you identify in your data set. In the traffic accident example, the date on which the accident occurred can serve this purpose.

Deviation measures the average change in a facet over *time*. Content Analytics first establishes the norm for the facet. In the traffic accident example, the norm is the average number of daily traffic accidents. When severe weather occurs, such as an ice storm, the roads might be icy, and an unusually high number of accidents might occur in a particular region during that time. Deviation calculated by Content Analytics shows a higher frequency in that time line relative to data for other time lines and highlights the data during that time. Other data is not highlighted in response to over-time changes. The deviation aims to measure how facets deviate from the average frequency over a specific time period.

1.4 Content Analytics architecture

This section provides details about the overall architecture of Content Analytics including a description of each component, the flow between components, scalability, and security.

1.4.1 Main components

Content Analytics consists of the following major components as shown in Figure 1-2:

Crawlers Extract content from enterprise data sources.

Document processors Process crawled documents in preparation for indexing.

Indexer Builds a document index for high-speed text mining and analysis.

Search run time Services client search and analytic requests.

Text miner application Used to perform text analysis.

Administration console Used to configure and administer Content Analytics.

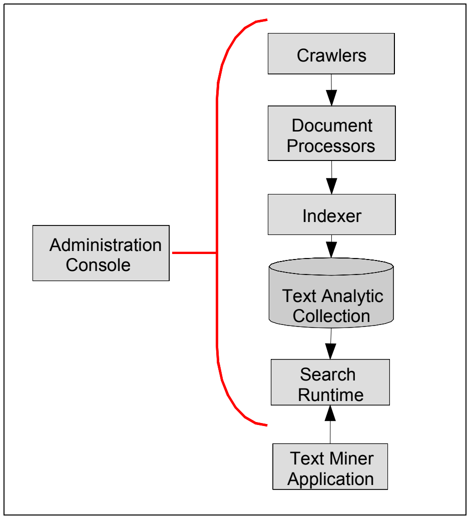


Figure 1-2 Component architecture

Crawlers

Crawlers extract content from the various enterprise data sources at intervals configured by the administrator. Crawlers are available for many different types of enterprise data sources. Content Analytics supports the following categories of crawlers:

- Web-based crawlers that support the HTTP/HTTPS and Network News Transfer Protocol (NNTP)
- Enterprise data source crawlers that support IBM Content Manager, IBM FileNet® Content Manager, and IBM Lotus Web Content Management

- Collaboration crawlers that support IBM Domino®
- File systems crawlers
- WebSphere portal crawlers
- Relational database crawlers
- Email crawlers that support IBM Lotus Notes® and Microsoft® Exchange

Most crawlers can crawl multiple data sources of the same type and do so with multiple threads, the number of which are configurable.

You can set up rules for crawlers to govern their behavior. For example, you can specify rules to control how a crawler uses system resources. The set of data sources that is eligible to be crawled constitutes the *crawl space*. You can edit the properties of a crawler to alter how it collects data. You can also edit the crawl space to change the crawler schedule, add new sources, or remove sources that are not to be searched any longer.

Crawlers can be started and stopped manually, or schedules can be set up. When scheduling a crawler, you specify when it must run initially and how often it must visit the data sources to crawl new and changed documents.

Some crawlers, such as those for web and NNTP sources, run continuously. After specifying the URLs or NNTP news groups to be crawled, the crawler returns periodically to check for data that is new and changed. The frequency of crawling the web and NNTP data sources is determined automatically based on configuration guidelines that specify the lower and upper limit of crawl frequencies. For example, you might specify a lower limit of weekly crawl frequency and an upper limit of daily crawl frequency. These limits guide the determination of the actual crawl frequency with statistics accumulated about the percentage of changes detected over past crawl frequencies.

Each data source type is associated with a different crawler type. For example, all IBM DB2 data sources have a DB2 crawler, and file system data sources have a file system crawler. However, multiple crawlers can be defined for different data sources of the same data source type. For example, one crawler can be defined for a set of DB2 tables in the human resources system, while another crawler can be defined for a set of DB2 tables in a data warehousing system.

One or more security tokens can be associated with crawled documents. A security token plug-in can be written to generate relevant security tokens for a document by using appropriate lookups of access control lists (ACLs).

Document processors

The *document processor* component is responsible for processing crawled documents and preparing them for indexing. In this component, the various text analytics are applied to the crawled documents. Each text analytic is referred to as an *annotator* because its job (in most cases) is to annotate a document with additional information that it extracts or deduces from the document content.

Annotators are built according to the Unstructured Information Management Architecture (UIMA) standard. UIMA is an open source standard sponsored by the Apache organization.[1] Content Analytics is UIMA-compliant and has a UIMA document processing pipeline built in. With this pipeline, you can plug in a prescribed number of text analytics annotators to process and extract what you need from your text (as illustrated in Figure 1-3).

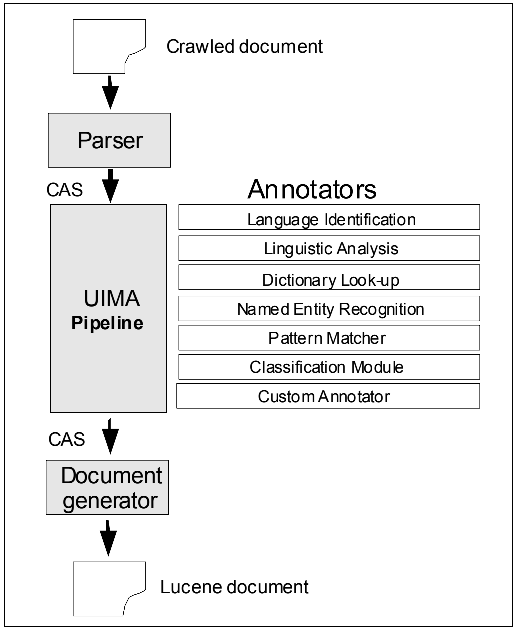


Figure 1-3 Document processor architecture in Content Analytics

In the UIMA pipeline, Content Analytics provides a set of required annotators that are necessary to begin the text analytics processing:

Language Identification annotator
Identifies the language of each document.

Linguistic Analysis annotator
Applies linguistic analysis for each document.

[1] See http://uima.apache.org/

These annotators cannot be changed. Content Analytics also comes with the following annotators that you can configure or enable next in the UIMA pipeline, each adding their own annotations to the incoming documents:

Dictionary Lookup annotator

Matches words and synonyms from a dictionary with words in your text. The annotator also associates the keywords with user-defined facets.

Named Entity Recognition annotator

Extracts person names, locations, and company names. This annotator can only be enabled or disabled. It cannot be modified.

Pattern Matcher annotator

Identifies patterns in your text by using the rules that you defined. The annotator also associates the patterns with user-defined facets.

Classification Module annotator

Classifies content into categories.

You can also add your own custom annotator to the pipeline, which is run at the end.

For more information about annotators, see the following chapters:

- For the Dictionary Lookup and Pattern Matcher annotators, see Chapter 7, “Performing content analysis” on page 279.
- For the Classification Module annotator, see Chapter 9, “Content analysis with IBM Classification Module” on page 357.
- For all other annotators, see Chapter 11, “Configuring annotators” on page 449.

UIMA and its constituent annotators use a Common Analysis Structure (CAS) to represent each document as it is processed. The CAS allows for the independent development of text analytic annotators. Each annotator adds its own annotations to the CAS. The CAS in its entirety is then made available to the next annotator in the UIMA pipeline.

The parser subcomponent is responsible for converting the crawled document in its native format into a CAS. The output of the UIMA pipeline is a CAS that contains the original document content plus any annotations added by the annotators. The document generator reads the CAS information and prepares the content for indexing by converting the document to a Lucene document.

Indexer

The *indexer* component is responsible for building a highly optimized index of document content that is suitable for high-speed text mining and analysis. The index is based on the open source *Apache Lucene indexer* with IBM extensions. IBM is an authorized committer of the Lucene open source project and frequently contributes selected features and functions back to the Lucene community.

When started, the indexer automatically indexes documents after they are processed by the document processors. You can manually perform a full rebuild of an index at anytime. This option is useful if you add a text analytic to the UIMA pipeline and you want to include its results in your text analytics collection. If you have enabled the document cache option for your collection, it is not necessary to recrawl your documents again. In this situation, the document content is obtained from the document cache, which is the temporary repository of crawled content.

Search run time

The search runtime component is a server-based component that is responsible for servicing client search and analytic requests. Client service requests are made by using the SIAPI. SIAPI is a programming interface based on Java that operates remotely by using the HTTP/HTTPS protocol. The search and text miner applications are example SIAPI client applications that make service requests to a search runtime component. (See “Using the SIAPI” on page 41 for more information about SIAPI.)

A single text analytics collection is associated with at least one search run time or multiple search run times to support large-scale multiuser environments. The search runtime components are not dependent on the indexer component and are designed for continuous operation. To maintain this independence, they operate on copies of their associated text analytics collection. For more information about this topic, see 1.4.2, “Data flow” on page 21.

Text miner application

The text miner application is the component from which text analysis is performed. The text miner user interface is browser-based and communicates with the text miner web application that runs under either Jetty or WebSphere Application Server. Jetty is the default installation. The text miner web application issues SIAPI client requests to the search run time associated with a given text analytics collection. The search runtime server component can either be installed locally or remotely from the text miner application.

With the text miner application, you can easily switch between multiple text analytics collections within a given browser session but can operate on only one text analytics collection at a time. Concurrent but independent analysis of multiple text analytics collections can be achieved with multiple browser sessions, one for each text analytics collection.

For more information about the text miner application, see Chapter 5, “Text miner application: Basic features” on page 143, and Chapter 7, “Performing content analysis” on page 279.

Administration console

Content Analytics comes with a robust administrative component. With this component, you can create and administer collections, start and stop components, and monitor system activity and log files. You can also configure administrative users, associate search and text miner applications with collections, and specify information to enforce security. Similar to the text miner application, the administration console is browser-based.

For more information about the administration console, see 4.2, “Administering Content Analytics” on page 84.

1.4.2 Data flow

The primary entity in a Content Analytics system is the text analytics collection. A *text analytics collection* is an optimized index of your document content that is designed for high-speed text mining and content analysis.

The administrator’s job is to create, configure, and manage text analytics collections (see Figure 1-2 on page 16). Business and research analysts use the text miner application to analyze their data in the text analytics collection.

After a text analytics collection is initially created, the administrator configures one or more crawlers for the collection. Crawlers are responsible for extracting the data from the enterprise and storing the data as documents in a cache on disk. Crawlers can be scheduled on a recurring basis or started on demand and operate independently of the rest of the system components.

The documents are read from the cache by the document processors. Document processors run all configured text analysis annotators against the content and prepare the content for indexing by formatting the content into a Lucene document (see Figure 1-3 on page 18). Because document processing is a prerequisite step before indexing, the document processor component is started when the indexing subsystem is started.

The indexing component stores the Lucene documents into the text analytics collection. The flow of documents just described is typically a continuous operation when the indexing component is active. As soon as a crawler extracts documents from the enterprise and places them into the cache, the document processors pick them up and prepare them for indexing. Likewise, after a Lucene document is available to be indexed, the indexing component picks it up, indexes the document, and stores the information in the index. This series of steps continues until all of the designated documents are fetched by the crawlers and stored into the index. The next time the crawlers run, the steps are repeated.

Let us assume that the crawlers have done their work and a text analytics collection has been successfully built. Let us also assume that you have added another text analysis annotator to the document processing pipeline to identify and extract person names. In this case, it is logical to assume that a full recrawl of your documents is necessary so that they enter the document processing pipeline and are eventually reindexed with person names added. But if you have enabled the document cache option for your collection, it is not necessary to recrawl your documents again. In this situation, the document content is obtained from the document cache, which is the temporary repository of the crawled content.

After a text analytics collection is built, the collection is copied to its corresponding search runtime components. With their own copy of the text analytics collection, a search runtime component can operate independently of the rest of the system components. When all of these components are installed on the same server, a copy of the generated index is not made with the search run time accessing the indexer generated collection.

For information about building the collection, see Chapter 4, “Installing and configuring IBM Content Analytics” on page 71.

Export points in the data flow

As documents flow through the system, they go through three major transformations. After each transformation point, the documents can be optionally exported for consumption by other external applications. In this way, Content Analytics is used as a text analytics platform by an application using only those parts that are useful to the application.

Figure 1-4 on page 23 illustrates the three export points within the Content Analytics data flow. For all three export points, the exported documents can be exported to either the file system in XML format, directly into a relational database, or to a CSV formatted file. You can write a custom Java export plug-in to change the format and destination of the exported content. For example, you can develop a custom plug-in to feed the documents directly into a case management system.

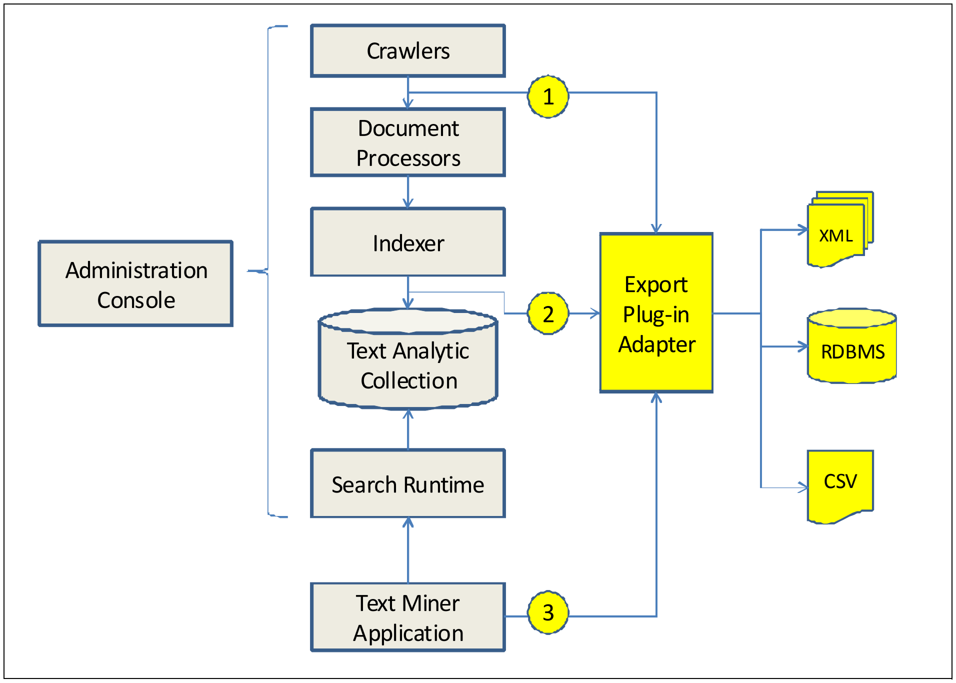


Figure 1-4 Export points in Content Analytics

The first export point is after documents are crawled and stored in the document cache. At this point, an application can intercept the original document in its binary form and any additional metadata provided by the crawler.

The second export point is after text analysis by the document processors and indexing. Here you get the same content available after the first export point plus any annotations added by the text analysis annotators in the UIMA pipeline.

The third and last export point is after a search has been performed. At anytime during data analysis in the text miner application, you can export the current search results set. Here the same content is available from second export point but filtered down to only those documents that match your current query.

For more information about export, see Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387.

Deep inspection

Deep inspection is a feature in Content Analytics that is a variation of the third export option mentioned in “Export points in the data flow” on page 22. After deep inspection has been enabled for your collection using the administration console, a deep inspection icon is displayed in each text miner view (except for the documents view). When you click the deep inspection icon, a batch submission is made to deep inspection for subsequent background processing of your current query. The results of the deep inspection are saved into an XML or CSV file on the file system. The exported results from deep inspection are not the individual search result document but rather the currently selected keywords along with their frequency counts and correlation values. In this way, deep inspection provides the same results as viewed from the text miner application.

The primary reason for using the deep inspection function is when the number of keywords to be analyzed is quite large (for example, greater than 500) and it is possibly causing a noticeable delay in the normal operation of the text miner application. The text miner application is designed as an on-demand text mining tool that supports rapid calculations in response to changes in a query. As the number of keywords increases, the time it takes to perform the calculations also increases. The default is 100 keywords. When increasing the value, be aware of the implication on the performance.

Deep inspection runs in the background and saves its results as an XML or CSV file in the file system. You cannot view the deep inspection results in the text miner user interface.

For more information about deep inspection, see 10.7, “Deep inspection” on page 431.

1.4.3 Scalability

Content Analytics is an advanced text analytics product that can grow with your needs. As you start to realize the benefit of text analytics, you might find new applications for the product resulting in more text analytics collections and business analysts using the text miner application.

Adding more users to the system can consume the resources of the search runtime component, eventually compromising the performance of the text miner application. To improve performance, you can use Content Analytics to scale the search runtime component across multiple machines. After each text analytics collection is built on the indexing server, it is copied to each of the search runtime servers. A single copy can be shared between search run times if using a shared mass storage device.

The document processing component, with its applied text analytics, can also be a compute-intensive and time-consuming task. Content Analytics supports scaling of the document processing component across multiple machines. Documents are distributed across the pool of available document processor machines on a round-robin basis.

Additional document processor and search runtime machines can be added to the system in real time on an as-needed basis. You are not required to shut down the system and restart it to use the added servers. Also any given document processor server can be switched to serve as a search runtime machine and vice versa.

Fore more information about scalability, see 15.6, “Scalability” on page 595.

1.4.4 Security

A text analytics collection contains the content extracted from the enterprise. Therefore, a system must provide stringent safeguards to protect content from unauthorized access. Content Analytics addresses this need in several ways:

- Administrative access control
- Collection level access control
- User application role-based access control
- Data encryption

Security is an advanced topic that is beyond the scope of this chapter. For more information about Content Analytics and security, see Appendix A, “Security in IBM Content Analytics” on page 633.

2

Application design and preparation

To achieve the most from your textual content, it is important to spend some time thinking about the overall design of your text analytics application and your expectations of the resulting analysis. You must do this step before attempting to work with IBM Content Analytics. Skipping this step might result in data mining results that you did not intend to have or might not find useful. A carefully designed text analytics collection and corresponding text miner application can unlock the value in your text and become an indispensable component of your business intelligence assets.

This chapter addresses text mining application design and data preparation. It includes the following sections:

- Use-case scenarios
- Data considerations
- Design guide for building a text analytics collection
- Programming interfaces

2.1 Use-case scenarios

Content Analytics is a powerful tool that can be used to provide invaluable insight from your text. The insight depends on the questions you ask of your data and that you must consider in the context of your particular use-case scenario. This section presents common scenarios in which Content Analytics has been used to provide actionable results and insight. From these scenarios, you can better understand where Content Analytics can be of help to you, and thus help you to design and prepare for your text miner application.

2.1.1 Call center

Call centers are a common and often a necessary component of a customer service department for a company. Generally, through the call center, a customer contacts a company representative to ask a question or resolve an issue with a product or service offered by the company. Call center agents log conversations between them and their customers. These logs are free-format text records (textual data). Content Analytics can be used to mine this data to find insight into the types of questions that customers ask and discover unexpected correlations between multiple products mentioned in the call center records.

Valuable information can be locked inside these call center records. This information can indicate the overall sentiment of the customer and their interaction with the company. For example, the text can reveal if the customers are generally satisfied with their purchase or service, or more importantly, if they are dissatisfied. Equipped with this information, you can make better decisions about what is working and what is not. If something is not working, the transcript might reveal the root cause of the problem, such as poor services. See 3.3.1, “Voice of customer” on page 60, for more detailed information about this use-case scenario.

2.1.2 Insurance fraud

Insurance fraud is a serious crime that affects all of us in the form of higher insurance premiums. Insurance fraud is the act of requesting reimbursement for expenses incurred because of an insured accident, procedure, or service that did not actually happen. Insurance companies spend enormous amounts of time and effort to detect insurance fraud. By reducing the amount of fraud incurred, insurance companies can increase their profit and be more competitive through lower premiums.

Insurance fraud analysis has traditionally been performed with only structured information about the insurance claim forms and other supporting documents.

Such information includes the name of the claimant, their date of birth, and the date of the last insurance claim. With Content Analytics, the information in the text, such as notes made by the insurance adjuster or comments made by eye witnesses, can now be used. With Content Analytics, new patterns can be exposed from the data extracted from this text. For example, the text might reveal, in many fraudulent cases, a lack of trauma in the claim report. The reason is that it is unlikely that people might harm themselves to satisfy a claim. Lack of trauma can then become one factor among many others that indicates potential fraud.

2.1.3 Quality assurance

Good quality is a measure of the commitment of a company to its customers to deliver products that they can depend on. A reduction in quality can result in a loss of customers to your competition. Therefore, it is important to maintain a high level of quality assurance by monitoring the continued quality of your products. Manufacturers have long recognized this important aspect of their business and put in place policies and procedures that track certain quality markers in their products.

For example, auto manufactures use the maintenance and repair records from their dealerships as one source of information to track the quality of their automobiles. Maintenance records typically contain structured information coded by the technician that identify the particular repair that was made (for example, a transmission repair). Usually a comment field accompanies the report in which the technician enters a more detailed description of the problem and the repair solution.

In this textual information, Content Analytics can provide an early warning of a potential quality issue that is emerging for a particular model of car. For example, Content Analytics can reveal that several technicians from different dealerships cited faulty wiring harnesses due to premature corrosion and that this particular problem only occurred for a particular model of car. With this information, you can detect the problem earlier and take corrective action before it becomes a more expensive problem.

2.1.4 Content assessment

Because of dramatic reductions in server and disk storage costs, companies are experiencing explosive growth in their digital content. Their digital content has shifted from being maintained at a central computing center to being highly distributed throughout the company on servers connected through an intranet.

Because of this growth and distribution of content, companies have found that keeping track of what is available and what is important is increasingly difficult. The information that is critical to the business of a company might be overlooked in some cases and then not placed under proper business controls and management. Similarly sensitive information might be recorded in documents that the company is unaware exist. These documents can be a potential liability that is used in legal actions for or against the company. Therefore, companies must identify these documents and place them under proper document management control.

Content Analytics can help you analyze your content to determine what is important and what is not. Through its advanced text analytics, Content Analytics helps you to see how your digital content is naturally organized.

For example, you can see how documents are clustered together by the frequency and collocations of the words used in those documents. You get an idea about what is in your documents without actually reading them. After you isolate a set of documents using Content Analytics, you can instruct Content Analytics to retrieve those documents and migrate them directly into a document management or a records management system for subsequent control and management.

In addition to Content Analytics, IBM has bundled two other products to provide a comprehensive content assessment solution. The bundle is referred to as *IBM Content Assessment* and includes the bundling of the following products that are designed to work together as described earlier:

- Content Analytics 2.1
  - Crawls and analyzes content from various enterprise content sources
  - Allows the exploration of content from these various sources
  - Supports the export of subsets of documents for additional investigation
- IBM Classification Module
  - Integrates into Content Analytics to provide document categorization and clustering
  - Integrates into Content Collector to enhance workflow routing of documents
- IBM Content Collector
  - Imports documents exported from Content Analytics into an IBM Enterprise Content Management (ECM) repository

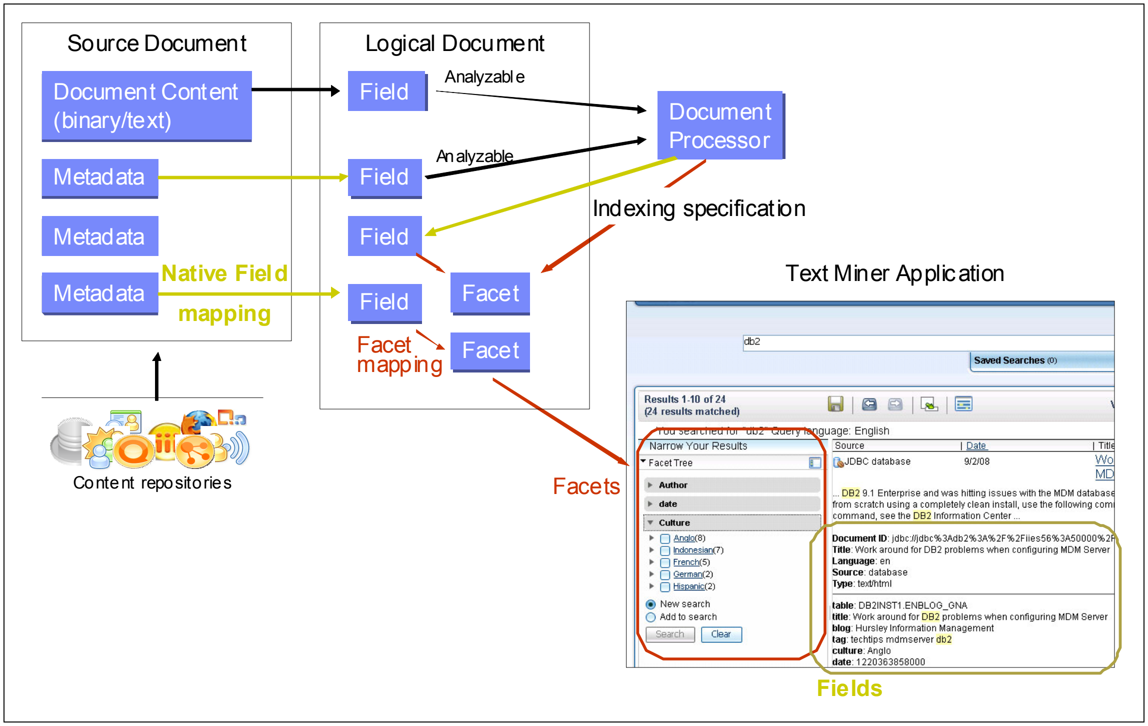
For more information, see Chapter 12, “IBM Content Assessment scenario” on page 483.

2.2 Data considerations

This section presents several aspects of data that you must consider when designing your text analytics application. Understanding these data characteristics and how they interact with the Content Analytics data model is important.

2.2.1 Content Analytics data model

Figure 2-1 illustrates the Content Analytics data model.



JDBC database 9/2/08

... DB2 9.1 Enterprise and was hitting issues with the MDM database from scratch using a completely clean install, use the following com command, see the DB2 Information Center ...

Document ID: jdbc://jdbc%3Adb2%3A%2F%2Fiies56%3A50000%2F
Title: Work around for DB2 problems when configuring MDM Server
Language: en
Source: database
Type: text/html

table: DB2INST1.ENBLOG_GNA
title: Work around for DB2 problems when configuring MDM Server
blog: Hursley Information Management
tag: techtips mdmserver db2
culture: Anglo
date: 1220363858000

Figure 2-1 Content Analytics data model

A single text analytics collection consists of one or more documents crawled from one or more content repositories illustrated on the left in Figure 2-1. A crawled document consists of its main content (often referred to as the *document body* stored in binary or text form) and one or more metadata fields. These fields are often referred to as *native fields* because they originate from the source document.

A logical document shown in the middle of Figure 2-1 on page 31 is created from the original document and from any information added by text analytic annotators during the document processing phase. Only one logical document can be created for a given collection. A logical document consists of fields and facets.

Fields (also known as *search fields*) are populated with data obtained either from the original source of the document or from information added by annotators in the Unstructured Information Management Architecture (UIMA) pipeline. Fields identified as analyzable have their content passed to the document processor for text analysis by the UIMA annotators. In turn, the annotators can populate new fields and facets in the logical document with information they have derived from the text.

A set of default fields is already defined for your text analytics collection. Default fields are common across all crawler source types, such as the document identifier or document title. Crawlers are aware of these fields and populate them with data from the document automatically. Besides the document identifier, title, and date field, it is not mandatory that you use the default fields. You can ignore them or delete them from your list of defined search fields.

Facets are similar to fields but are used to represent the different dimensions (or views) across all documents in your collection. Similar to fields, facets are populated with data from the original documents or from annotators in the UIMA pipeline. In the text miner application, facets are displayed in the Facet Navigation pane on the left side of the Text Miner window (at right in Figure 2-1 on page 31). However, fields are shown with each document search result in the middle of the Text Miner window. Default fields are differentiated from any custom fields by a visible horizontal divider line in the search result. Default fields are listed first above the divider line.

Search field: A search field refers to a field that exists within your collection and can be acted upon. A search field can be used for search, returned in a search result, used to sort a search result, and much more. A search field in a Content Analytics collection is semantically similar to a column in a database table.

2.2.2 Structured and unstructured sources

Content Analytics supports a broad spectrum of content sources, with more than 25 different types of enterprise sources, including everything from plain text files to relational databases. The amount of structure in the data from these supported sources varies widely and must be considered when designing your text analytics application.

Highly unstructured data sources, such as text files and web pages, consist mostly of text and have few structured fields or metadata associated with them. In this case, you must rely more heavily on the various text analytics in Content Analytics to extract meaningful information from the text portion of the content.

Some data sources can have an almost equal amount of structured and unstructured data. Documents crawled from document management systems fall into this category. For example, documents maintained in a Domino database can have a rich set of structured fields and textual fields. The contents of a facet can be populated by the values in your structured fields and the values extracted from your unstructured data using text analytics.

Data sources, such as relational databases, have a high degree of structure. Relational databases support a near infinite set of relationships that can be defined for the data. Relational databases can also contain text in VCHAR, CLOB, or BLOB fields. You normally have these kinds of fields analyzed by the text analytic annotators.

The data model for Content Analytics is that of a simple document where a document consists of a set of fields. Your job is to map more complex relationships (defined in your relational database for example) to the document model used in Content Analytics. This task might require preprocessing of your data outside the Content Analytics product.

2.2.3 Multiple data sources

For a given text analytics collection, you can define one or more crawlers to acquire the documents to be analyzed. If your documents are from a single source, you configure a single crawler for that particular source type. Crawler configuration includes mapping the native fields in the source document to the search fields you defined for your text analytics collection. If you have a single document source and hence one crawler, this mapping step is trivial to perform. Most, if not all, of your search fields are derived from the native fields in the crawled documents.

If your text analytics collection obtains data from multiple sources, you must be aware of the ramifications. A single common logical document is created and indexed in the text analytics collection for each document retrieved by a crawler. Documents retrieved by different crawlers *are not* merged into one logical document that is then indexed into your text analytics collection. Consequently, it is possible to have a logical document with sparsely populated search fields depending on the diversity between crawled data sources.

This functionality can work to your advantage depending on your data. For example, suppose a regulatory agency wants to monitor the performance of medical devices in the marketplace. The agency developed a form that captures all of the pertinent information when a medical device experiences a failure. The forms can be submitted by three different sources: hospitals, patients, and the manufacturer of the device. In addition, suppose that the three sources use different formats and destinations for submitting the form data. In this situation, you can create and configure three different crawlers, one for each source. Fortunately the data is consistent across each crawler. Each form retrieved by each crawler extracts the same set of fields and maps them to their corresponding search field.

Now consider a case where the data (fields) retrieved by multiple crawlers are different. For example, suppose you have insurance claim forms that you want to analyze. You define and configure a crawler to retrieve these forms. The majority of the information that you need is in the claim form except for the age of the claimant, which you deemed important in your analysis. However, the age of the claimant is in a relational database. Therefore, you configure a relational database crawler mapping the age column in the relational database to the age search field.

The result is two separate documents in the collection for each claim:

- One for the insurance claim form itself where most of the search fields are populated with values from the form.
- One that represents the age data extracted from the relational database. The document contains only one populated field value (age). All of the remaining search fields in this document are empty.

This behavior might not be what you want. When you require a join of information from multiple data sources, you must perform your own extract, transform, and load (ETL) processing outside of Content Analytics to merge the data into one logical document. After the data is merged, Content Analytics can then use a crawler to crawl these merged documents into a text analytics collection.

2.2.4 Date-sensitive data

An important feature of Content Analytics is its ability to identify trends and patterns in your data that occur over time. In order for Content Analytics to perform this task, it must base its calculations on one or more date values that are consistently contained in each logical document of the text analytics collection. Without these date fields, Content Analytics cannot perform any of the date-sensitive calculations, which in turn are used to calculate the time series, deviations, and trends views of the text miner application.

At a minimum, you must identify one or more date fields for a given collection in order for the time series, deviations, and trends views to be operable. When dealing with a single data source, the task can be as trivial as identifying which date fields to use. You can also have the date fields in your logical document populated by multiple crawlers of different types. Use care to ensure that the date fields from these sources are semantically the same and represent the same information.

2.2.5 Extracting information from textual data

For the textual data in documents, you can use the annotators that are provided by Content Analytics to extract useful information.

You can configure the following key annotators for text analysis:

- Dictionary Lookup annotator. Match words and synonyms from a dictionary with words in your text. The annotator also associates the keywords with user-defined facets. For configuration information, see 7.3, “Configuring the Dictionary Lookup annotator” on page 299.
- Pattern Matcher annotator. Identify patterns in your text by using the rules that you defined. The annotator also associates the patterns with user-defined facets. For configuration information, see 7.4, “Configuring the Pattern Matcher annotator” on page 309.

In addition to these annotators, you can categorize your documents by using IBM Classification Module (a separate product). Classification Module is a powerful classifier that you can train by presenting it with sample text examples for each category that you want to recognize. After the Classification Module is trained, you can activate its corresponding annotator in Content Analytics to automatically categorize your documents as they pass through the UIMA document processing pipeline. To use document classification, you must purchase a separate IBM Classification Module license.

The IBM Classification Module can also be used to group your documents into self-organizing clusters based on the frequency of words and their collocation to others words used in your documents. This type of text analytic provides insight as to the main topics being discussed in your documents, with each cluster representing a potential main topic. At anytime, you can activate document clustering for a text analytics collection by using the administration console. The use of document clustering does not require the purchase of an IBM Classification Module.

The IBM Classification Module can also be used to support concept-based searching. Concept-based searching can return relevant search results without necessarily containing any of the keywords used in your search expression. For example, the search expression “home damage caused by bug infestations” might return documents about termite damage even though the term “termite” was not used in your search expression. This a powerful way to search. To use concept-based searching, it is assumed that you have already obtained the IBM Classification Module product and have trained a knowledge base using a sample set of clustered results for your collection.

For configuration and usage information, see Chapter 9, “Content analysis with IBM Classification Module” on page 357. For other annotators that are available in Content Analytics, see Chapter 11, “Configuring annotators” on page 449.

2.2.6 The number of collections to use

A single installation of Content Analytics supports the creation and maintenance of multiple text analytics collections. Each text analytics collection consists of its own defined crawlers, text analytics, and indexing options. Each collection operates independently of the other text analytics collections. Multiple text analytics collections are useful when you want to use Content Analytics for different and independent research efforts.

One or more text analytics collections can be organized into a grouping that is assigned an *application ID* by the Content Analytics administrator. Through the use of application server roles, users can be associated with a particular application ID. In this fashion, users can be restricted to which text analytics collections they can access through the text miner application. For more information, see Appendix A, “Security in IBM Content Analytics” on page 633.

Even though multiple text analytics collections can be created, keep all required documents for a specific analysis effort in a single text analytics collection. The reason for this approach is because the statistics calculated in the text miner application are designed to apply to only one collection. They are not designed to be calculated across multiple text analytics collections.

2.3 Design guide for building a text analytics collection

This section explains the steps to build a text analytics collection. The text analytics collection is the central component to your research. The collection supports the rapid search and discovery features provided in the text miner application.

2.3.1 Building a text analytics collection

Use the following steps to build a text analytics collection:

1. Design your search fields for the text analytics collection following the guidelines in 2.2, “Data considerations” on page 31, and 4.3.1, “Designing a sample collection” on page 87.
2. Create your text analytics collection as explained in “Creating the collection” on page 90.
3. Create the search fields for your text analytics collection as explained in “Creating the search fields” on page 96 and “Mapping the native field to the search field” on page 103.
4. Create your facets and map your search fields to the facets as explained in “Creating facets and mapping search fields to facets” on page 106.
5. Define your crawlers and map the source native fields to your search fields as explained in 4.3.3, “Defining and configuring a crawler” on page 123.
6. Run your crawlers to extract the documents from the enterprise as explained in “Starting the crawler component” on page 129.
7. Build your text analytics collection as explained in 4.3.4, “Building an index in the text analytics collection” on page 132.
8. Verify your text analytics collection as explained in 4.4, “Verifying that the collection is available” on page 138.
9. Repeat steps 2 through 8 to make adjustments to your text analytics collection as explained in 2.3.3, “Planning for iteration” on page 40.

A text analytics collection is compiled from data extracted from various content sources in your organization. These sources include the file system, relational databases, email systems, the web, and document management systems.

2.3.2 A walk through the building process

When designing your text analytics collection (step 1), you first determine which native fields in the source repositories to extract and index into your text analytics collection. You also identify the search fields that will be populated with data extracted by the text analytics in the UIMA pipeline. The resulting list of search fields represents the superset of fields that are available to your text mining application.

For each search field, you define how the application must use it:

- Returnable

  The content of the field is returned with the search results. Returnable is the default setting. There can be times when you might want a field to be searchable but not returned for display such as a salary field. Also, a large number of returned fields in the search results (for example, a hundred or more) can impede performance. Choose these fields sparingly.

- Free text search

  This type of field can be searched with a keyword portion of a search expression. Normally the body of a document (for example an email) is searched in this manner. You can also specify fields, such as the title field, to be examined (that is free text searched). If you select the free text search field attribute for a particular field, optionally you can specify whether to use the contents of the field in the makeup of the dynamic summary that is displayed for the search result. The dynamic summary highlights the words that match the keywords specified in your query.

- Fielded search

  This type of field can be explicitly referenced in a search expression. For example, you might want to find all documents where the author field contains "John Doe" expressed as `author:"John Doe"`. If you select the fielded search attribute, you also can specify whether an exact match is required (that is all the words in the field must match) and whether the match is case sensitive.

- Parametric search

  This type of field is of type numeric or date and allows you to use algebraic expressions such as less than, equal, greater than, in-between, and combinations thereof during searches.

- Sortable

  This type of field indicates whether it can be sorted in the search results. By default, your search results are sorted by relevance and date.

- Analyzable

  The content of this type of field can be analyzed by Content Analytics. Content Analytics applies advanced text analytics to extract parts of speech, phrase constituents, named entities, and any other annotators that you have configured for the content of this field.

- Faceted field

  This type of field is displayed in the Facet Navigation pane in the left pane in both the text miner application and the search application. A faceted field typically has a finite set of discrete values that can be enumerated and is *not* normally a free text field.

After you design your application fields and facets, you are ready to create the text analytics collection by using the administration console (step 2 on page 37). Be sure to indicate that collection is a text analytics collection rather than a search collection.

After the collection is created, you configure the search fields (step 3 on page 37). You perform most of the work for defining your search fields in step 1 on page 37. In step 3 on page 37, you manually enter them into the system. If you have a large number of search fields, you might find it more convenient to import the definitions by using a single XML file.

In step 4 on page 37, you define the facet hierarchy to be used in the text miner application. The facet names that you assign to the hierarchy are for display purposes only. The search fields that you map to the display facet names are used to populate the tree.

In step 5 on page 37, you define the crawlers for the target sources. Crawlers extract documents from the back-end sources and make them available for indexing. For each crawler, you map the native fields in the source documents to the indexed fields in the text analytics collection. The field names do not need to match. It is common to encounter fields from different back-end sources that are semantically the same but have different field names. For example, the from field in an email document can be mapped to the author field in the index, However, in an office document, you map the creator field of the document to the same author search field.

> **Order of the steps:** The order of these steps can vary slightly depending on your preference. For example, after you create the text analytics collection in step 2 on page 37, you can define your crawlers next for the collection (step 5 on page 37). However, you cannot map your native crawled fields to the search fields of the collection until you define the search fields in step 3 on page 37. The system does not prevent this sequence of steps. At anytime, you can go back to previously defined crawlers and add or change the mapping of its native fields to the search fields of the collection.

In step 6 on page 37, after you create the crawlers, you run them to extract the needed documents from your enterprise data source. If this is the first time for running your crawlers, limit the number of documents crawled to a smaller, more manageable set. After you are confident with the configuration and building of your text analytics collection, you can run the crawlers to retrieve the entire corpus of documents.

Content Analytics offers continuous operation. If all components are running (for example, crawlers, indexer, and search run time), as soon as crawlers retrieve documents, they are processed and indexed. After being indexed, the documents are made available to the text miner application for analysis. Again, if this is the first time you use the product, step through each component independently. Start and stop them with the completion of each task. Remember the number of documents that have been crawled up to this point.

In step 7 on page 37, build the text analytics collection. Follow the instructions in 4.3.4, “Building an index in the text analytics collection” on page 132. With each click of the **Refresh** button on the administration console, you see the progress of the index build and the number of documents currently in the index. Depending on how many documents were crawled and how many text analytics you have configured, building and completing your text analytics collection can take some time. You can get a rough idea of the remaining time by monitoring how long it takes to process a certain amount of documents per minute. Then apply that factor to the remaining documents that will be crawled. You are then informed by an appropriate message when the collection has been completed.

By step 8 on page 37, you are ready to start the search runtime function for collection. Use the text miner application to verify your configuration.

2.3.3 Planning for iteration

You are most likely to find yourself iterating through the steps in the design guide in 2.3.1, “Building a text analytics collection” on page 37. Start with a small set of sample data to rapidly validate your design assumptions. Depending on the application, this data set can range any where from a few hundred documents to tens of thousands of documents. The more documents you have, the longer it takes to crawl, parse, analyze, and build the text analytics collection. Select a number that is reasonable for your needs.

Based on our experience, we always started with a few hundred documents to verify our field mappings, the text analytics applied, and facets. With fewer documents, we did not get meaningful statistics such as correlation values. With each iteration, we increased the number of documents by a factor of ten to double check that the text analytics are doing what we expect. With a larger data set, we were more likely to encounter occurrences of specific entities and patterns in the text that our text analytics were looking for.

After you are certain that everything is working as expected, you can use Content Analytics on your entire corpus of documents. The length of time to process your entire corpus depends on the hardware being used and the number of documents in your corpus. You can get a rough estimate of the time to complete this task based on your previous iterations through the smaller sets of data.

Keep in mind the following considerations when iterating through the building of your final text analytics collection:

- If you discover an error in the mapping of native fields to search fields (for example, you miss a field), rebuild index can fix the field mapping.
- If you change the search field attributes or change the facet tree, it is not necessary to recrawl all of your data. In this case, you simply redeploy the index and then rebuild the index only.

2.4 Programming interfaces

Content Analytics offers two kinds of programming interfaces from which to build search, text analytic, and associated administrative applications. Each kind is described in this section.

2.4.1 Search and Index API

Content Analytics comes with a robust and flexible Search and Index API (SIAPI). In the Content Analytics implementation of SIAPI, the search server can be accessed remotely. The SIAPI is used to develop the text miner application. You can use it to build your own text analytics application or to create applications that administer collections.

This section provides general guidelines when designing and using the SIAPI for your custom text analytic application.

Using the SIAPI

To use the Content Analytics SIAPI, follow these steps:

1. Obtain an SIAPI implementation.
2. Obtain a SearchService object.
3. Obtain a Searchable object.
4. Issue queries.
5. Process the query results.

Obtaining a SIAPI implementation

Begin writing an SIAPI-based application by obtaining an implementation factory object. SIAPI is a factory-based Java API. All the objects that are used in your search application are either created by calling SIAPI object-factory methods or are returned by calling methods of factory-generated objects. Therefore, you can easily switch between SIAPI implementations by loading different factories. The SIAPI implementation is provided by the `com.ibm.es.api.search.RemoteSearchFactory` class.

Obtaining a SearchService object

Use the factory object to obtain a SearchService object. With the SearchService object, you can access searchable text analytics collections on the Content Analytics search server. The SearchService object must be configured with the host name and port of the Content Analytics search server and optionally with the required locale for receiving error messages. Configuration parameters are set in a java.util.Properties file. The parameters are then passed to the `getSearchService` factory method that generates the SearchService object. Call the `getAvailableSearchables` method to obtain all of the Searchables object that are available for your application.

Obtaining a Searchable object

Use the SearchService object to obtain a Searchable object. A Searchable object is associated with a text analytics collection on the search server. With a Searchable object, you can execute queries and get information about the associated collection. Each Content Analytics collection has an ID. When you request a Searchable object, you must identify your application by using an application ID. Contact your Content Analytics administrator for the appropriate application ID.

If the Content Analytics search server is configured with global security turned on, you need to provide a password. The password is used to authenticate your application.

Issuing queries

When you obtain a Searchable object, you can issue one or more queries against that Searchable object. To issue a query against the Searchable object, follow these steps:

1. Create a Query object.
2. Customize the Query object.
3. Submit the Query object to the Searchable object.
4. Get the query results, encapsulated in a ResultSet object.

The SIAPI uses an extensive query syntax to process queries.

Processing query results

With the ResultSet and Result interfaces, you can access query results. The SIAPI has various methods for interacting with the ResultSet and individual Result objects.

More information

For additional information about using the SIAPI, see the general programming guide, *Content Analytics Programmer Guide*, SC19-2874. Also look in the following directories:

- SIAPI sample programs in the `installation directory/samples/siapi` directory
- The SIAPI Javadoc information in the `installation directory\docs\api\siapi` directory

2.4.2 REST API

Content Analytics also provides a Representation State Transfer (REST)-based API for the development of search, text analytic, and associated administration applications. In particular, you can use the REST API to perform the following tasks:

- Manage collections.
- Control and monitor components.
- Add documents to a collection.
- Search a collection and federated collections.
- Search and browse facets.

The REST API offers the following benefits:

- A language independent and pure remote call. Any client modules are not required to use the API.
- Easy to understand. Almost all communications between client and server are in text format, and you can use your web browser to try the API.
- Integration in Web 2.0. Any clients that support HTTP can use the API. You can build client applications on various platforms.

The REST API has two categories: the Search REST API and the Admin REST API. The Search REST API is available on machines that have a Content Analytics search server component deployed. It is available on any machine that has a search role. The search server is listening on the search server port (port 8394 by default).

The Admin REST API is available on machines that have the master server role. The master server port is 8390 by default. These port numbers can be changed during product installation.

Common rules for using REST API

You can use both the HTTP GET and HTTP POST methods to call most of the REST APIs. The HTTP POST method is recommended for reasons of greater security.

Parameters

The REST API has the following parameters:

- XML and JSON are available as return types. You can add `"&output=application/xml"` to get a return value in XML and `"&output=application/json"` to get a return value in JSON.
- All parameter values must be URL escaped.
- If a parameter allows you to specify multiple values, you can specify a string separated by the vertical bar (|), for example `name=value1|value2|value3`. If the vertical bar is used in a value, it must be escaped by using a backslash (\) and the escape character must be escaped by using a double backslash (\\). For example, `xyz=abc\|de\\f` is parsed as `key:xyz` and `value:abc|de\f`.
- Some parameters require a JSON formatted string. The JSON formatted string can be specified as follows:

```
          <path for API>?name={"innerName1":"value1",
"innerName2":"value2"}&anotherName=value...
```

Error response

The error responses are standard HTTP error response messages. Error messages are being mapped to the appropriate HTTP response codes. More detailed error information is appended to the HTTP response in a format that matches the content type and language, for example, that is requested.

The response format might differ depending on the value specified in the HTTP request for the output parameter. The HTTP response header Content-type might indicate the message format returned.

More information

For additional information about using the REST API, see the general programming guide, *Content Analytics Programmer Guide*, SC19-2874. Also look in the following directories:

- Sample programs in the *`installation directory`*`/samples/rest` directory
- Javadoc information in the *`installation directory`*`\docs\api\rest` directory

3

Understanding content analysis

This chapter provides details about the process of content analysis and how IBM Content Analytics can be used as a tool to help you analyze large amounts of textual data. From this content analysis, you can gain actionable insight from your data. To be successful, you do more than use the product and perform a series of mechanical operations. You must have a deep understanding of what Content Analytics does and how it works. To take full advantage of Content Analytics, you must understand what you can analyze, what you can expect, and how you can interpret and use its output.

This chapter includes the following sections:

- Basic concepts of Content Analytics
- Typical cycle of analysis with Content Analytics
- Successful use cases

3.1 Basic concepts of Content Analytics

Textual data can be complex and ambiguous. Because of the ambiguities of natural language, textual data often obscures factual information and insight that you can otherwise use and act on to make better business decisions. This type of information is difficult to understand and process by using automated methods. Consequently, businesses are handicapped without considering this large body of information.

Content Analytics specifically unleashes the value trapped in your textual data. It is a tool for reporting statistics and to obtain *actionable insights*, which are business insights that lead to actions for better business operations. Content Analytics is used to reduce your manual workload of text analysis and to enable a higher level of analysis that provides insights not previously attainable.

In our experience, Content Analytics users reach greater levels of awareness and achievement with their data. By using Content Analytics, users can take actions based on the insights they obtain from their data and make their business operations more efficient and better managed.

To gain the insights of this textual data, new users must understand the basics of content analysis.

3.1.1 Manual versus automated analysis

When manually analyzing textual data, an analyst typically classifies the data according to a predefined set of classifications, tallies the number of documents that conform to each classification, and then reports on their distributions. The problem comes when the amount of data to be analyzed is large. An analyst or a group of analysts can find it increasingly difficult to process and discover anything out of the ordinary that is not predefined.

For example, assume that you have over 1,000,000 survey forms from people indicating their plans for the upcoming weekend. Each survey entry consists of information such as the person’s age, profession, and gender, and description of their weekend plans. The information in this survey can be valuable to businesses because it enables the businesses to predict demands for goods and services from the respondents and to subsequently prepare for them. However, because the weekend activities are described as free-form text, it is impossible to manually read through and analyze each of the 1,000,000 survey forms.

Consequently, a human analyst might randomly select a much more manageable subset of survey forms (for example, 1,000 forms or fewer). After reading a portion of these forms, the analyst might then define an arbitrary set of classes for the weekend plans (for example, shopping and traveling). Finally, the analyst might tally each of the survey forms according to the defined classes and produce the following result:

- 250 shopping
- 200 visiting friends
- 150 playing sports
- 100 traveling
- 300 others

Figure 3-1 shows the results in a graphical form.

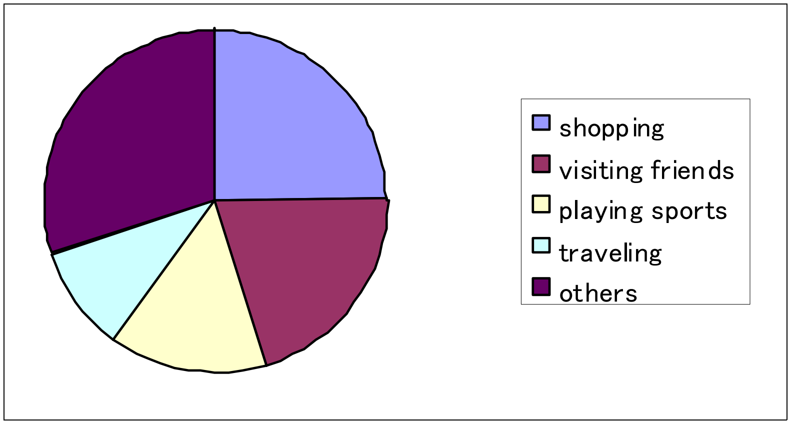


Figure 3-1 Distribution of weekend activities

Manual text analysis that follows this pattern can be a difficult and time consuming task to achieve. Defining a proper set of classes for the data is not always trivial. For example, the statement “Traveling to Florida to play golf with my friends” might belong to multiple classes, such as traveling and sports. The statement “I'm planning to go hiking unless it rains. Otherwise, I'll be shopping.” is ambiguous as whether to classify it as sports or shopping depending on the weather condition.

The composition of the classes can be difficult. What must a classification be and to what level of detail? In our example, for the sports category, we want to further understand the kinds of sports (for example, tennis and hiking) that people are planning for the weekend. Originally, the analyst only classified the entries to the single high-level sports category. Drilling down into each sports activity after the first round of analysis is done requires additional workload to tally more survey forms that correspond to each specific sporting activity.

With Content Analytics, you can now use an automated means to analyze your textual data. In our example, Content Analytics can easily process 1,000,000 survey forms. By defining appropriate facets for various viewpoints, the survey data set might lead to valuable insights for businesses.

For example, the tourist industry might define a travel destination facet that consists of major places for travel and an activity facet that consists of typical activities performed during travel. With these facets, the tourist industry can now analyze the type (such as age, profession, and gender) of people who tend to travel to specific areas of the country for certain kinds of activities. By defining a shopping place facet and a purchase item facet, retailers can analyze the type of people who tend to buy what at where. This information can now be extracted from your textual data.

The distribution of high-level classifications tends to be similar over time and is not useful. The role of Content Analytics is to change the process of generating distribution reports into recommendations for action. The essential goal of the analyst must be changed from document classifier and chart maker to interpreter of the analytical results, identification of actionable insights, and planning for action.

For more information about defining facets and performing analysis, see Chapter 5, “Text miner application: Basic features” on page 143.

Challenges in text analytics

Natural language contains many ambiguities that are difficult for automated methods to resolve. For example, the phrase “Time flies like an arrow” can be interpreted in several ways. The word *like* might be a verb meaning “to be fond of” or an adjective meaning “similar to.” Likewise the word *flies* might be interpreted as the verb “to fly” or a plural noun as in “multiple fire flies.” Polysemous words (words with different meanings) are difficult for automation. For example, the name *Arizona* can be a place or the name of a person. The word *saw* in the sentence “They saw a girl with a telescope” can be interpreted as the past tense of the word “see” or an object that you use to cut something. Besides “with a telescope” might modify either “saw” or “a girl.”

In addition to words being ambiguous, the concepts conveyed by the text can also be ambiguous. This ambiguity is demonstrated in our example while trying to classify the surveys on weekend plans. It is difficult to classify the hiking activity as either a sports activity or a travel activity. The result is often subjective and can differ from person to person. Even simple errors, such as misspellings in the original text, can cause problems in its proper interpretation.

Because of these challenges, the analytical results might not always match your expectations. These difficulties can affect the overall distribution of items with associated frequencies being suspect. Even the order might not be reliable because modifications made to the dictionaries can inadvertently change the order. For example, if we define CD as a synonym of “compact disc,” the number of records that contain the words “compact disc” might increase erroneously because some of them really refer to “certificate of deposit.”

Having accurate frequency counts as much as possible is preferred. Content Analytics provides several solutions for improving the accuracy in its text analytics. It uses tools such as the dictionary and pattern matching rules components (Dictionary Lookup and Pattern Matcher annotators). However, improving the overall accuracy in your text analytics can be an elusive task. It is much more productive instead to spend more time in the analysis phase and action taking phase of your work. With Content Analytics, you can examine *all* of your data and that the large amount of data in itself leads to the power of the analysis and value of the result.

Using human intuition and understanding during manual analysis is the best way to improve accuracy but quickly becomes impractical as the amount of textual data increases. As the amount of data increases, human involvement needs to shift from the manual reading and understanding of text to the automated analysis of the corpus as a whole. You cannot add more people and expect the overall accuracy to improve because different analysts produce different results. Even the same analyst can produce different results over time. However, Content Analytics can display distributions from various viewpoints over the whole of the data. By using Content Analytics, the criterion remains the same for the whole corpus of documents.

3.1.2 Frequency versus deviation

The terms *frequency* and *deviation*, as defined in 1.3, “Important concepts and terminology” on page 9, are important to understand in the context of content analysis.

Frequency is the number of documents that contain keywords identified by the text analytics. Frequency might not be reliable because of the difficulty encountered by the text analytics as mentioned in “Challenges in text analytics” on page 48. Even if you can reach 100% perfection in text analytics, the numbers produced might still not reflect reality.

For example, suppose that you want to analyze customer contact records that involve a specific product failure. In this case, you want to know the total number of units sold that are experiencing the failure or the total number of people who have encountered the product failure. It is not practical to expect that all of the customers might report the failure. Only some of them might call the customer contact center. If the failure is serious, many people might report the problem. If the failure is not serious, only a few people might call in. In the customer contact center, the agents might inadvertently leave out critical information when recording the problem as described by the customers.

Because of all of these factors, even after you make a significant effort to improve the accuracy of your text analytics, you cannot determine the true number of customers who experienced the problem nor the number of products that caused the problem. Figure 3-2 illustrates this point.

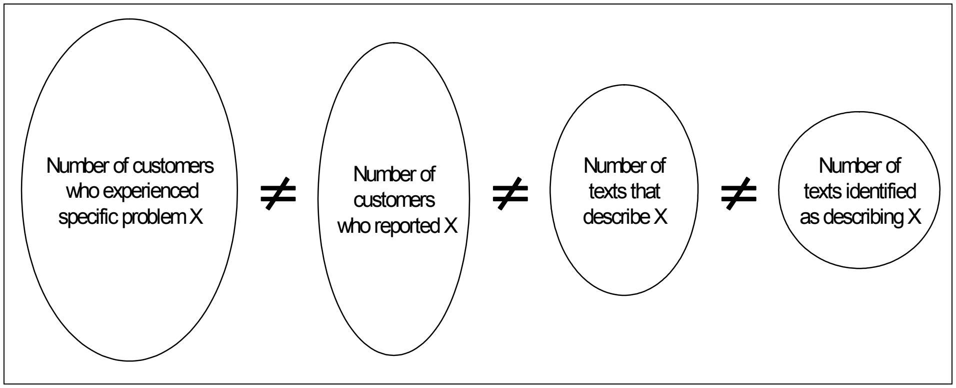

Number of customers who experienced specific problem X
≠
Number of customers who reported X
≠
Number of texts that describe X
≠
Number of texts identified as describing X

Figure 3-2 Frequency of textual items not matching reality

Even though the frequency counts might not be as reliable as you want, their deviations often lead to valuable insights. For example, 10% of the calls this month on product A were associated with problem X, and last month only 3% of the calls on product A were associated with problem X. In this case, the change or *deviation* is an indicator of a potential problem that is worth further investigation. That is assuming that no surge in selling product A in recent months nor more demand in using product A in recent months has occurred.

Likewise, consider an example where 10% of the calls this month on product A are associated with problem X, where only 3% of calls this month on product B are associated with problem X. In this case, it is better to take action and discover why product A is having a higher incidence rate than product B (assuming that functionality and quality of product A is similar to product B). This example illustrates how the use of Content Analytics can you help obtain actionable insights.

Often legitimate reasons can explain the changes and deviations. Changes and deviations are typically based on a real-life phenomenon related to your products or customers. This situation often provides a great opportunity to improve the business by dealing with product failures early or coping with customer behaviors before customers become disenchanted and leave. Although frequency numbers provide some insight, it is more important to focus on the deviations and changes in those numbers to gain greater actionable insight.

In the example of the weekend activities survey, say that you identify about 10% of the survey forms that mention sports as an activity. Consider that the text contains a description of a sport, such as tennis. It might not indicate that the respondent actually plans to play tennis over the weekend but rather plans to watch a tennis game or buy tennis shoes over the weekend. The Content Analytics-based analysis can capture non-sports activities as sports activities if Content Analytics is configured naively. The real power of Content Analytics is to analyze the entire corpus of data and treat the mistakes as *noise* when you focus on changes and deviations.

With Content Analytics and its broader analysis of all your data, you can gain insight and discover new relationships in your data. For example, you might discover that teenagers are more active in sports than older people. Within sports activities, golf might be strongly correlated with people in Florida, and hockey might be strongly correlated with people in Minnesota. Although some of these correlations might seem obvious, the nonobvious correlations are important to watch for and are the ones that Content Analytics highlights.

3.1.3 Precision versus recall

When dealing with a large amount of textual data, you must consider *precision* (accuracy) and *recall* (coverage) in the analytical results. Precision is the ratio of the correctly returned results as compared to the total returned results from Content Analytics. You can think of the precision of Content Analytics results as the number of true positives. Recalls refers to the ratio of the correctly returned results as compared to the total number of correct results in the data set.

For example, from the 1,000,000 survey forms about people's weekend activities, you might look for data that describes dining at a French restaurant. One thousand documents are returned from Content Analytics that contain the words "French restaurant" and "eat." However, such results usually contain documents that do not describe dining at a French restaurant, but contain such phrases as "We will make our anniversary plan to eat at a French restaurant," "We will eat at a hamburger shop next to the French restaurant," and so on. If the total number of documents that correctly describe dining at a French restaurant is 700 out of the 1,000 returned documents, the precision (of the results) is 70% (`700/1000`). That is 70% of the returned results from Content Analytics are correct.

The approach of getting results with the query "`French restaurant`" and "`eat`" might not capture all the data in which people describe dining at a French restaurant. The activity of dining at a French restaurant might be described as "`we are planning to have dinner at a French restaurant`," and the name of a specific French restaurant might be used instead of the literal word "French restaurant."

Thus, in the overall survey, data might exist that describes the activity of dining at a French restaurant without using either the words “eat” or “French restaurant.” The correct number of documents that describe the activity of dining at a French restaurant probably exists outside the 1,000 returned results that you received from the system. Consider that the total number of correct results (from the entire data set) is 2,000, and you received 700 (correctly returned documents from the 1,000 returned results) that describe the activity of dining at a French restaurant. The recall (of the result) is 35% (`700/2000`). That is, you captured 35% of all the correct results in the entire data set.

Precision and recall are often incompatible. If you aim for a higher recall because you do not want to miss any relevant documents, you might be faced with too many irrelevant documents (noise) because of lower precision. If you aim for higher precision because you do not want to be faced with irrelevant documents, you can miss relevant documents because of a lower recall.

In general, aiming for higher recall is often more difficult and time consuming than aiming for higher precision. To achieve higher precision, the basic process entails adding constraints in your query to eliminate the noise. To aim for a higher recall, you must craft increasingly complex query expressions to search and capture all relevant documents.

To analyze trends and characteristics of the data with Content Analytics, focus on high precision data. As long as the data set is large enough, you can obtain enough samples for meaningful statistical analysis even when the recall is low. On the contrary, if the precision is low, the data might have too much noise, and the trends and characteristics identified with such noisy data might not be reliable.

Special applications, such as those that are aimed to identify critical problems related to safety or legal compliance, might require a higher recall. For such applications, Content Analytics provides powerful functionality to gain clues for achieving higher recall.

To achieve higher recall value, you must first capture high precision data by adding constraints for patterns and eliminate noise. After you achieve a target precision value, look for relevant expressions in the data. Then you must use the relevant expressions that are identified from the second step to capture more correct data in the entire data set.

For example, to capture data with the activity of dining at a French restaurant, you might be able to achieve a higher precision value with the pattern of “`eat`,” without “`plan`” or negation in its front. It might immediately be followed by “`at`” and “`French restaurant`,” except for the determiners between “`at`” and “`French`.” If the precision is reasonable, you can use the Facets view to identify expressions that are relevant to the activity of dining at a French restaurant such as wine, cheese, sommelier, baguette bread, and foie gras. By analyzing expressions relevant to such expressions with the Facets view of Content Analytics, you can identify expressions to capture more recall of your data. In this case, you can capture the activity of dining at a French restaurant and include alternative expressions for French restaurant such as "`place for French cuisine`" and names of French restaurants.

3.2 Typical cycle of analysis with Content Analytics

In 2.3, "Design guide for building a text analytics collection" on page 36, you go through the administrative steps required to build a text analytics collection. In this section, you revisit those steps in the context of the overall content analysis process described in this chapter. The result is content analysis of your data that meets your expectations and allows you to gain actionable insight to make better business decisions.

Using Content Analytics consists of iterating through the following steps:

1. Set the objectives for the analysis.
2. Gather the data.
3. Analyze the data.
4. Take actions based on the analysis.
5. Validate the effect.

The data analysis step (step 3) consists of iterations of multiple steps as follows:

1. Apply text analytics and generate the index.
2. Perform analysis using the text miner application.
3. Generate or modify dictionaries, patterns, or both.
4. Reapply the text analytics and regenerate the index.
5. Repeat data analysis by using the text miner application while verifying changes in each facet.
6. Regenerate and modify dictionaries and patterns as necessary.

3.2.1 Setting the objectives of the analysis

Your content analysis objectives depend on what you want to do and what the data can reveal. For example, if you want to decrease product failures, early identification of product failures is a reasonable objective. However, if none the data describes product failures, it is not feasible to set early identification of product failures as the objective.

You cannot always determine if your objectives are reasonable until you perform the actual content analysis. However, it is important to consider several objectives and list them up front as potential objectives. Even though some objectives are not feasible because of a lack of measurable data, you can improve the data for such objectives in the next cycle.

3.2.2 Gathering data

Data is an essential part of your analysis. For example, if you want to understand how your customer perceives your product or services (positively, negatively, or neutral), you need to gather data that contains opinions expressed by your customers.

As you go through each iteration of your analysis, think about how to improve your data for better results. Sometimes this task requires you to gather or integrate additional data, and sometimes it means you must cleanse the data for input.

For example, you might analyze the sentiment of customers recorded in a call center. In this case, data improvement can consist of integrating email messages from customers for a broader coverage of data. You can also ask the call center agents to get additional information from customers such as the reason they chose the product. This process might also involve modifying crawlers or data input systems.

3.2.3 Analyzing data

After you set your objectives and gather the data, you can start analyzing the data. The text miner application is your primary tool for content analysis. The process of analysis is an iterative one. With each cycle, you refine your dictionary, matching rules, and input data, so that you can discover new insights, perform necessary actions, and achieve the objectives you set.

Analysis with the text miner application

At the beginning of the content analysis cycle, Content Analytics runs text analytics with a set of default dictionaries and patterns that captures grammatical information such as nouns, verbs, adjectives, adverbs, and other parts of speech. This information can be useful for the initial analysis of your data. A review of the Part of Speech facets gives you a synopsis as to what subjects are involved and an overview of the major concepts in the data set.

You must also browse through other facets in the Facets view to understand the information that is available for analysis. Facets populated from structured data are important at this phase of analysis because the data is generally unambiguous and is much more informative compared to your textual information at this time. Always explore other views for facets that might be interesting. Use your curiosity and imagination.

Text miner views summary: You can browse through different views of the text miner application to discover insight. See Chapter 6, “Text miner application: Views” on page 217, for a detailed description of the views. For your convenience, a brief summary of each view is provided here:

Documents view Shows a list of documents that match your query. The view shows the actual content of individual documents and their metadata. By reading the original text, you can quickly verify that the text is supporting the reported results. This view helps to determine if additional fine-tuning is needed to further improve the accuracy of the results.

Facets view Shows a list of keywords for a selected facet. Each keyword has a corresponding frequency count and correlation value. This view is useful for seeing the keywords that make up a given facet in your data.

Time Series view Shows the frequency change over time. This view is used to analyze frequency and to select a range of documents for analysis for a given time period.

Trends view Shows sharp and unexpected increases in frequency over time. Usually a sharp increase warrants further investigation. If the highlighted trend is associated with an undesirable trait, for example product failures, then you must investigate how to reverse the trend.

Deviations view Shows deviation of keywords for a given time period. This view is focused on how much the frequency of a facet deviates from the expected average for a given time period (not from its past history as in the Trends view). You use this view to observe seasonal patterns in your data or patterns that occur on a monthly or weekly basis.

Connections view Shows the correlation of keywords from two selected facets in a graphical way. This view enables you to visually see the connections between to selected facets.

Dashboard view Shows a configured dashboard layout with one or more graphs and tables in a single view. With this view, you can see multiple views (that are of your interest) at the same time.

Figure 3-3 shows a collection of all the text miner views.

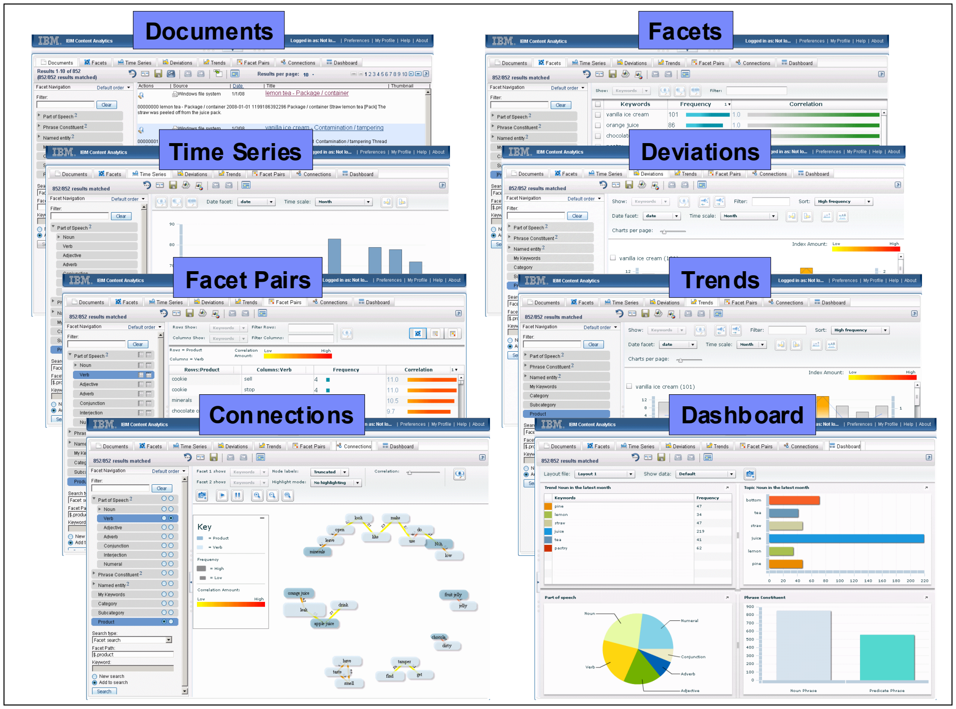


Figure 3-3 Summary of the text miner application views

Each view in the text miner application provides its own unique kind of insight from your content analysis. Analysis of distributions based on day of week and month of year in the *Deviations view* often leads to interesting insights. You might be able to find unexpected patterns of weekend activities and seasonal differences. With the *Trends view*, you can also obtain actionable insights. If something undesirable is increasing, you might need to reverse the trend. The *Facet Pairs view* unveils remarkable associations if you select the appropriate pairs of facets. Repeated selections of different facets to compare is crucial to your success. Content Analytics is designed to be interactive and scalable for repeated iterations through your data.

To understand the content of the data, it is essential to read the original text with the *Documents view*. Some people misunderstand *text mining* as something that allows you to avoid reading the actual documents. Even though text mining allows you to avoid reading the entire amount of textual data, it still requires you to read selected portions of text for better understanding. After you identify interesting keywords or patterns, make sure to focus on the data with the specific keyword or pattern to verify the fact. What is thought to be an interesting pattern might be caused by noise in the data, errors in the text analytics, or duplication of data. Content Analytics is designed to make it easy for you to verify your results with highlighted keywords in your original text. You can usually get to the original text with a couple of clicks.

By using the text miner application during your content analysis, you might identify and select new objectives. It is important to be flexible so that you can set some initial objectives suitable to your data.

Generation and modification of dictionaries and patterns

After you set the objectives of your content analysis, the next step is to tune your text analytics to capture the right information that matches your objectives. You can usually extract most of the information you need by using the Dictionary Lookup and Pattern Matcher annotators.

A dictionary consists of a list of words and phrases supplied by you. In the dictionary, you create facets that describe different aspects of your data that you want to investigate and analyze. For example, facets can be colors and shapes. You associate each facet with a list of words (also called *keywords*) that you create. For example, you can create the keywords `yellow`, `blue`, and `red` and associate them with the Color facets. You can also create the keywords `square`, `circle`, `triangle`, and `rectangle` as keywords and associate them with the Shape facet. When documents are processed in Content Analytics, the text in each document is broken into individual words (or phrases) and then checked against the dictionaries that you built. If a match occurs, that document is associated with the facet, and the frequency count of the facet is incremented by one.

Refining your dictionary with synonyms is one way to fine-tune your text analytics. For example, by registering multiple terms for “International Business Machines” as synonyms of each other (in this example, IBM) you can treat them as a single token. With the pattern matcher, you can identify relationships with parts of speech patterns. For example, you can add a pattern to extract a noun immediately followed by the verb “love” for extracting the relationship of something being loved.

The essential purpose of the dictionary and pattern matching definition is to build different viewpoints for your analysis that will meet your objectives. It consists of defining the appropriate facets and their corresponding expressions.

For example, if your objective is to identify early warnings of product failures in customer complaint records, you need to focus on defining facets relevant to product failures. Definitions of facets that are relevant to product failures optimize your analysis. Otherwise, the trends or patterns you want to analyze will be hidden in various irrelevant facets.

You must define facets in such a way that their expressions are semantically or grammatically correct. For example, when analyzing computer product failures, a logical facet to define is one that might identify specific components that failed in the product. You can label this facet the Failed Component facet. The values of this facet might consist of nouns such as “keyboard,” “mouse,” “display,” and “battery.” You can define another facet that identifies the cause or type of failures. You can label this facet the Failure Type facet. The facet can consist of verb patterns with diverse parts of speech, such as “crack” and “break,” and be used as a noun or a verb.

In general, start with the dictionaries first, rather than the patterns, especially when the facet consists of simple nouns. It is much easier to define entries in the dictionary than to define grammatical patterns. However, using grammatical patterns is much more flexible in extracting ambiguous terms. Most often you need to use the pattern matcher to handle the conjugation of verbs.

The best resource for words and synonyms to be defined in your dictionary is your textual data. The list of nouns in the Facets view (with the Noun facet underneath the Part of Speech facet) is generally the best candidate list. The list of compound nouns (in the Noun Sequence facet under the Noun Phrase facet, which is under the Phrase Constituent facet) also provides good candidates.

These facets represent how words, which are potentially important to your analysis, are expressed in your textual data. The official names of products are rarely expressed by default because they are not part of a normal set of generic nouns. In addition, abbreviations or short hand might be used in the textual data. Therefore, it is better to use the words as they occur in the text. The top values in these Parts of Speech facets are the most frequently occurring in the textual data. Starting with these words is efficient and effective because words with lower frequencies do not contribute as much to your deviation analysis.

After you identify initial expressions for a specific facet, it is often useful to apply correlation analysis for extracting more expressions for the facet. For example, in the PC help center example, we identify the word “`screen`” to be associated with the Component facet. One verb that highly correlated with the word “`screen`” in the data was the word “`turn`.” Among the nouns that highly correlated with the word “`turn`,” we found that the words “`monitor`” and “`display`” can be added as synonyms of the word “`screen`.” Also the words “`power`” and “`battery`” can be added to the Component facet.

For detailed procedures on generating and modifying the dictionary and patterns, see 7.3, “Configuring the Dictionary Lookup annotator” on page 299, and 7.4, “Configuring the Pattern Matcher annotator” on page 309.

The amount of effort it takes to develop your dictionary and patterns depends on your data and your objectives. Some users spend significant amounts of effort developing their dictionary and patterns and end up with a comprehensive and robust lexicon. Also many users have generated an initial dictionary within an hour or so of work and then modify it as they encounter more terms during their analysis. In either case, the use of dictionaries and patterns is essential for analysis and often has a significant impact on your analysis. It is important that you take the time to create an initial dictionary in support of your objectives for analysis.

Repeating the analysis with the text miner application

After you create or modify your dictionary and patterns, you must apply the text analytics again and regenerate the collection as described in 4.3.4, "Building an index in the text analytics collection" on page 132. After you reindex based on the latest dictionary and patterns, you are ready to analyze the results with the text miner application.

First, you must verify the changes in each facet. Some words and phrases might not have matched as you intended. Mistakes, such as misspellings and format errors, can cause inaccurate statistics for the facet. Expressions in the data might not always represent the concept you expected. The facets can represent a different concept altogether, and the concept you expected might be expressed differently. Thus, it is important to validate your dictionary and patterns after you generate or modify them.

In addition to the Facets view, use the Deviations, Trends, and Facet Pairs views while verifying the results of your dictionary and patterns. Even if the dictionary and patterns are in their early stages of development, analysis with the various views often inspires you to try better approaches for modifying the dictionary and patterns.

3.2.4 Taking action based on the analysis

The goal of your analysis is to take positive actions based on the insights acquired from your data. Because a large amount of textual data usually contains complex relationships full of various insights, it is ideal for action takers to be able to use Content Analytics during their decision making process. This way, they can consider various aspects of potential actions for better decisions.

The use of Content Analytics often leads to significant results when the analysts themselves can control the actions. Based on our experience, some companies have recognized the value of this new insight and have made the analyst team position a career path for executives.

3.2.5 Validating the effect

Evaluating the results of your actions is important for determining the validity and quality of your analysis and for planning of the next cycle of analysis.

Because the analysis of your data provides clues of insights unveiled from your data, you usually need to verify the discovered insights in the real world. For example, you might be analyzing complaint data about cars. By using Content Analytics, you can identify a rapid increase in the number of complaints that apply to a specific car model with a specific problem. With Content Analytics, you can also identify different facets that are strongly correlated to this problem, and such facets can indicate certain solutions and actions to take.

The increase in complaints might be caused by changes in the data input operation or a rumor about a potential problem (of which people might call in simply to inquire about the potential remored problem). Or it might be caused by a problem that exists in the database (for example, causing duplidates of the data). In any case, there must be a reason for the changes and deviations in the data, and it is usually worth further investigation.

Validation of the effect is important because it leads to improved analysis in the next cycle and setting of new objectives for the new cycle.

3.3 Successful use cases

Content Analytics provides infinite possibilities for your company by enabling you to take advantage of your data that is accessible to you. You inspire the outcome, which can be great insight or actions that enhance and improve your business operations, products, or services.

This section introduces successful use-case scenarios in which data is analyzed with Content Analytics. The intent is that these scenarios might inspire you to learn and use Content Analytics in innovative ways to enhance and improve your business. Each use-case scenario is based on real-life experience.

3.3.1 Voice of customer

Analysis of the voice of customer (VOC) has been a major application of Content Analytics. Analysis of VOC is critical for a business because it provides crucial information about customers and products. Regardless of how hard you test your products before shipment, unexpected use of, or a defect in, your products is unavoidable.

For example, a customer called into a PC help center and claimed that the cup holder on the PC was broken. Obviously, PCs do not have cup holders. After further conversation, the call center agent determined that the customer was actually using the CD tray as a cup holder.

The same VOC records can be used for other purposes. For example, one of the customer contact records in a PC help center contained the sentence, “`CX'S DOG ATE HER POWER SUPPLY`.” The VOC record further indicated that the agent looked up the part number for the specific power adapter and transferred the customer to the parts division. The development division can use this information to change the design or material of the power adapter because it might lead to a safety issue. Also, they might need to analyze similar cases with other animals and even young children. The PC help center can use this information to coach the agent to be more sensitive and comment on the condition of the dog in addition to helping to order a new power adapter. Such an attitude can impress customers and improve their satisfaction. The Sales division noted that this customer owns a dog, and that this person might be a potential customer for dog-related products.

The analysis of VOC can lead to cost reduction, improvement of customer satisfaction, and an increase in the hit ratio of target marketing.

Early identification of product failures

Early warning of product failures is one of the most promising applications of VOC analysis with Content Analytics. We have observed significant results across industries including manufacturing, catalog retailers, and service providers.

The PC help center estimated that, in the US alone, the savings from one of their early identifications of a specific product failure using Content Analytics was worth several million dollars. The PC help center received, on average, more than 10,000 calls per week from customers. The call center agents typed in a brief overview of each customer contact that recorded what each customer talked about and how the agent answered. Each call center record also contains structured information such as the machine type and a time stamp.

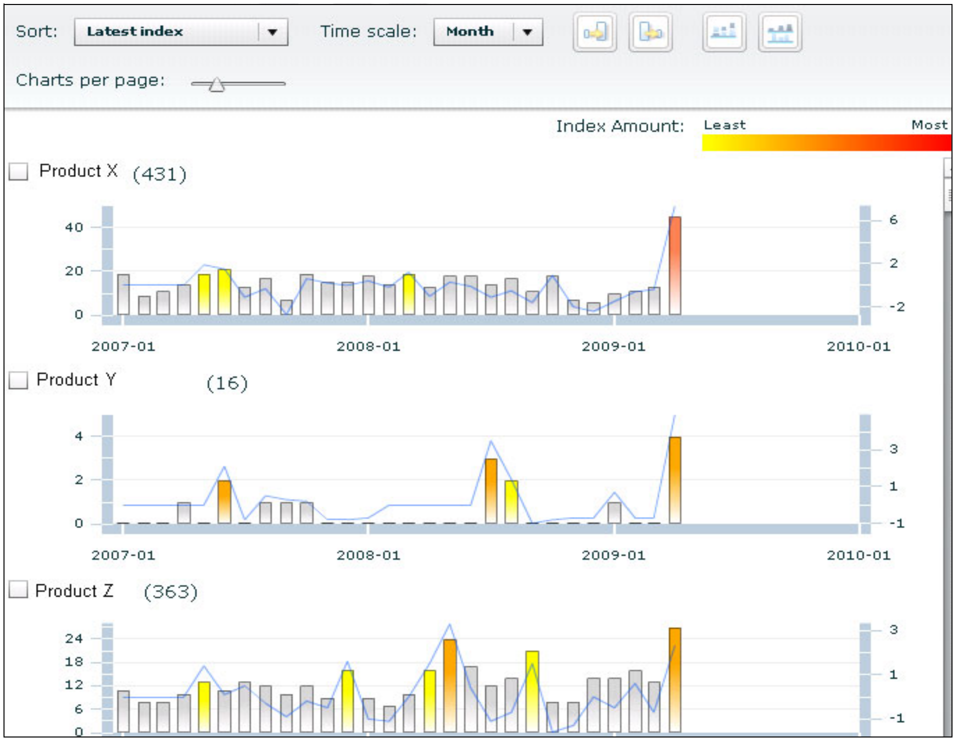
Before the introduction of Content Analytics, analysts in the PC help center manually analyzed approximately 300 call center records for a weekly report. With less than 3% of the call center records being analyzed, it was unlikely for the report to have a significant impact for the business. Also the task was expensive and required much effort to read through 300 records to prepare the report.

With Content Analytics, the PC help center used the entire data set, not just the 10,000 records for each week, but the millions of records recorded for the past couple of years. The PC help center also compared the data of the current week with the data from the past several weeks and compared the data of the same week with the data from the past couple of years.

The PC help center defined facets such as Product Name, Hardware, Software, Subcomponent, and Problem. A keywords distribution for product X that is significantly different from the keywords distribution in the same facet for comparable products can indicate a potential problem with product X. You can easily capture such differences (frequency and correlation) in the Facet Pairs view of Content Analytics. Consider the example where 5% of the records for a brand new computer contain the words “`LCD panel`” in the Hardware facet. In this example, only 1% of records for comparable computers contain the same words “`LCD panel`” in the Hardware facet. In this case, the Facet Pairs view in combination with the Hardware facet and the Product Name facet helps to identify the strong correlation of the new computer to the words “`LCD panel`.”

Analysis using the Trend view is also effective for early warning of product failures because such failures generally result in a rapid increase of customer calls. To do such analysis, you monitor the Trends view by setting the sort criteria to Latest Index. This sorting shows the products with the highest frequencies first in descending order. With this view option, you look at the top of the list in the view. The appropriate facet to be selected for this analysis is the Product Name facet to identify the product name with the most increasing numbers of calls. It is also useful for analyzing the subcomponents of a product or other attributes of a specific product after identifying a specific product for further investigation.

For example, Figure 3-4 shows that the number of records for product X increased the most for the latest month, April 2009, as compared to the other products.

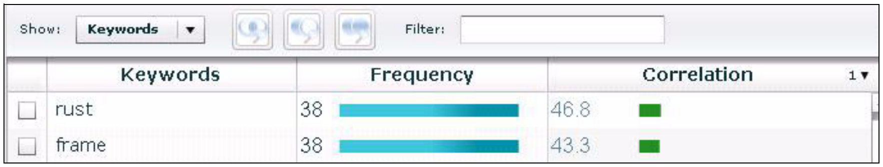


Sort: Latest index

Time scale: Month

Figure 3-4 Trends view: Identifying the highest frequency product in the latest month

By focusing on the data of product X for April 2009 and identifying significant words found in the data (Figure 3-5), you can conclude that product X might have product failures due to a frame rusting problem.



Show: Keywords

Filter:

	Keywords	Frequency	Correlation
□	rust	38	46.8
□	frame	38	43.3

Figure 3-5 Facets view: Identifying problem verbs with a high correlation

Early warning of product failures is effective and rewarding because it generally leads to actions that can be taken. Also it is often easy to evaluate the value of your early warning detection based on the expected number of product units sold and the cost of fixing the defect in each product with the failure.

Timely update of FAQs based on VOC

Content Analytics was also used at a PC help center in Japan for updating frequently asked questions (FAQ) based on customer contact records. The use of Content Analytics by the PC help center led them to achieve number one status in the problem solving ratio of web support among computer companies operating in Japan in 2003. Their web support also became number one in personal computer support ranking in 2004 according to a premiere PC magazine in Japan.

VOC is invariably the best resource for FAQ because it enables you to count the frequency of questions asked. However, relying on the frequency of questions asked is not always a reliable measure as discussed in “Challenges in text analytics” on page 48. In this scenario, the PC help center staff took a smart approach by focusing on the correlation values.

The staff identified high correlations for a specific type of product and its operation in comparison to similar types of products and their operations (using the Facets view and Facet Pairs view). Immediately, they added or modified entries of the FAQ database. In this way, they enriched the FAQ database with specific questions for specific products and their operations instead of general questions that might be more frequent in total. Because the solutions are targeted to specific questions, this activity significantly improved their problem solving ratio of web support and improved the overall customer satisfaction of web support.

As a result, access to web support increased while calls to the PC help center remained the same even though the number of products shipped were increasing. It allowed the staff to reduce the number of PC help center agents, which resulted in a cost reduction by several millions of dollars, with an increase in overall customer satisfaction for the PC.

The reason for the success of this operation was that the PC help center staff who made the analysis with Content Analytics took immediate action and updated the FAQ database.

Customer profiling for target marketing

Customer contact records usually contain valuable information about customers. Such information can be a valuable resource for product planning and marketing. For example, if dogs and cats are frequently mentioned in customer contact records, many customers might be interested in pet-related products.

One of the successful use cases of Content Analytics is to extract customer information from customer contact records for target marketing. Many companies keep track of the people and organizations who bought their products or who have expressed a strong interest in their products. The information is maintained in a relational database for customer relationship management and target marketing.

It is still uncommon to keep track of information about competitors’ products even though such information might be important to know. Many customer contact records in the sales department indicate the reason why they were unable to sell their products. For example, people or organizations they contacted might have bought competitive products or had different requirements. Such information is often described in customer contact records to justify their purchase decisions. Although this information can be regarded as useless if the sale failed, it can be valuable later as a customer conversion opportunity or for another division that sells different products or services. However, because there is no plan to use such information, agents usually describe them briefly in free-form textual data. This brief description makes it difficult for anyone to take advantage of such information.

By using Content Analytics, you can extract information about who is using what product and who is considering which solution. You can define product names and solution names in a dictionary or define patterns. For example, you can define a pattern that extracts nouns followed by the verb use.

A financial company in Japan took this approach to generate a target marketing list for various products. To sell government bonds, they looked for customers who mentioned (in their customer contact records) a preference for low risk. In order to sell foreign-currency-based products, the company looked for customers who mentioned international activities such as a business trip abroad. As a result, the company achieved a better hit ratio compared to their traditional approach of nearly random sampling. Selling agents were positive to this approach because most of the target customers showed interest in the product being sold, and the agents felt more comfortable when talking with these customers.

Human resource development

Customer contact records are also a good resource for improving agent skills. A customer contact center in Japan had a problem of maintaining highly skilled agents because the retention rate of the call center was low. Changes in senior agents affected the overall skill levels of the center and customer satisfaction. Because it is not easy to hire new agents with a strong information technology background, it was important to educate new agents effectively. After months of education based on FAQs, the new agents started taking calls and were educated through on-the-job training.

When a customer calls in and the first contact agent has trouble answering the customer’s question, the call center transfers the call to a second-level senior agent. Because the number of calls and the ratio of new agents were increasing, the number of transfers to the second-level agents was also rising too much.

The call center used Content Analytics to analyze customer contact records specifically for the following information:

- Calls transferred to second-level agents
- Calls taking a long time
- Calls based on product failures

The analysis of calls transferred to second-level agents allowed the call center to identify skill areas that required education. Through analysis of the calls that take a long time, the call center identified specific behaviors of each agent that led to long calls such as full verbal guidance of step-by-step operations instead of guidance to the web FAQ page.

Based on the insights gained through the Content Analytics analysis, the call center developed a more focused education course to improve the skill level of their agents. As a result, the skill level of their agents improved significantly in a short period. The number of calls transferred to second-level agents dropped to less than 20% from the number before the change in education. In addition, the call center forwarded their insights through the analysis of calls based on product failures to the product development team. This activity led to a remarkable reduction in the number of product failure calls.

People in this call center hold monthly meetings to report and share the new analytic approach used in Content Analytics. They also share any insights obtained through the new analytics. Thus, this call center has been running many cycles of analysis with Content Analytics over time, which has led to continuous improvement of this call center.

Best practices of sales agents

By analyzing customer contact records at an outbound call center for telemarketing, Content Analytics was used to identify the characteristics that separated good agents from mediocre ones. The agent enters a brief overview of each customer contact as a memo to prepare for the next call and a report to their managers. These customer contact records can be considered as sales activity reports.

As a part of this analysis, the managers of the agents were asked to select high performing agents. Then the differences in the words used between the reports of high performing agents and low performing agents were analyzed. However, such analysis did not lead to any valuable insights because the differences in the words used did not lead to meaningful actions. Therefore, the focused shifted to the best performing agent. Content Analytics revealed that this agent tended to initiate contact by expressing appreciation for some action the customer made such as visiting a workshop, submitting a survey result, or for using their product. This unique characteristic was confirmed in the Facets view where expressions for appreciation were highly correlated to this particular agent.

Another characteristic of this agent was that the person kept frequent contact with all of their customers, generally a couple of times each month, as was discovered in the Trends view.

Based on these findings, all agents were sorted by correlation to expressions for appreciation in the Facets view. As a result, agents were found to be divided into two groups, each at opposite ends. One group of high performing agents was highly correlated to expressions for appreciation. All members of this group showed similar patterns of frequent contacts with their customer according to the Trends view with customer names selected as the facet.

The other group of high performing agents was negatively correlated to expressions of appreciation. The contact pattern for this group, according to the Trends view, was different from the frequent contacts made by the other group of high performing agents. For almost all of their customers, they contacted them intensively during a relatively short time, spread across a couple of months to several times a month, with no contact before and after the intensive contacts.

Interestingly, other agents that the managers did not classify as high performers showed another contact pattern. In a sense, their contact pattern was not consistent. They did not contact all of their customers frequently nor intensively. They often contacted customers frequently for a while, and then after a couple of months of no contact, they reconnected.

The difference between the frequent contact group of high performing agents and the short but intensive contact group of high performing agents was analyzed, indicating a clear difference. The high performing agents in the intensive contact pattern have extremely high skill levels as revealed in the Facets view. The skill levels of the high performing agents in the frequent contact group were not necessarily high except for some communication-related skills.

This insight was used to motivate agents to acquire higher skills and to keep a frequent relationship with customers until they gained enough skills.

3.3.2 Analysis of other data

Content Analytics can be applied to various types of textual data other than customer contact records. However, the basic cycle of analysis remains the same for most types of textual data. You form your analysis by acquiring insights based on your objectives, perform actions based on the insights, and then validate your actions, which lead to better analysis in the next cycle.

For analysis of surveys, reviews, and bulletin board data, you can often apply approaches similar to the analysis of customer contact records because they also contain the sentiments of the customer. Through the analysis of such data, you can often identify product failures in their early stages. You can also analyze the positive and negative aspects of each product from the consumers' viewpoints, how customers chose each product, and how they are using them. If the data contains demographic information of the submitters, such as age and profession, you can analyze characteristics of opinions and behaviors based on generations and professions.

Project reports are another type of textual data that you can analyze for best practices and identify potential problems by using Content Analytics. You might be able to identify project in trouble or that are beginning to show signs of trouble by analyzing its textual data with Content Analytics.

Technical documents

By analyzing technical documents, such as patents, you can identify technical trends, the technical strength of companies, and so on. For example, by adding a technical term as a search condition, you can make a list of companies that filed patent documents that contain the term. With the Trends view, you can identify which companies filed, the number of related patents, and the time they filed them.

For example, if you type "`text mining`" as a query term, you can list the company names whose patents contain the phrase "`text mining`" in the Facets view. Then, by selecting the Trends view, you can see the competitive landscape of the companies working in text mining. By using the Facets view, you can identify terms relevant to text mining such as classification, query, and knowledge.

After you select a specific company from the Facets view, you can analyze technical terms relevant to the patents from that company. By analyzing technical terms of a specific company with the Trends view, you might be able to identify technical trends within that particular company.

3.4 Summary

Text mining is an interactive procedure with discovery throughout multiple cycles of analysis. The duration of a single cycle of analysis can vary greatly from days to months. In our experience, users often become accustomed to text mining operations for each new cycle, and the results lead to bigger impacts to the business as they become more experienced.

Large amounts of textual data often contain a great wealth of knowledge. With Content Analytics, you can acquire valuable insights depending on the objectives and viewpoints that you set.

Performing content analysis: If you already have a working environment that uses Content Analytics, have configured sample data, and are familiar with the user interface of the text miner application, jump to Chapter 7, “Performing content analysis” on page 279.

4

Installing and configuring IBM Content Analytics

The previous chapters introduce the IBM Content Analytics product. They explain its architecture and design, and provide considerations for collecting data for text analytics. This chapter explains how to install and configure Content Analytics. You see how to build a basic text analytics collection using the Content Analytics administration console with the sample data provided by the product. You also see how to verify that your system is installed and configured correctly.

This chapter includes the following sections:

- Installing Content Analytics
- Administering Content Analytics
- Configuring a text analytics collection
- Verifying that the collection is available
- Deploying the configuration

Tip: If you already have a working environment that uses Content Analytics, you can skip this chapter.

4.1 Installing Content Analytics

This section explains how to prepare for installing Content Analytics, such as determining the installation type and parameters. It also explains how to install Content Analytics step by step.

4.1.1 Process overview

Before you begin, you must understand the overall installation process. The general installation process consists of the following main tasks:

1. Plan for the installation.
2. Install the software.
3. Configure the software.
4. Perform verification of the installation.

Before you perform the system installation, review the latest product documentation.

4.1.2 Confirming the system requirements and supported data sources

Before you install Content Analytics, confirm the latest system requirements and supported data sources. Make sure that your environment meets the prerequisites identified at the following web addresses:

- For the latest system requirements:

  `http://www-01.ibm.com/support/docview.wss?&uid=swg27017944`

- For the latest supported data sources:

  `http://www-01.ibm.com/support/docview.wss?&uid=swg27017946`

Because the information might be updated from time to time, always review the latest information before you start.

Also, check for the latest IBM documentation on release and patch fix levels. Use the latest version levels for the platform you are installing.

See the following IBM web pages to help you find the appropriate information about Content Analytics and product support:

- Content Analytics Information Center

  http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp
- Content Analytics Version 2.2 information road map

  http://www-01.ibm.com/support/docview.wss?&uid=swg27017952
- IBM Support Portal for Content Analytics

  http://www-947.ibm.com/support/entry/portal/Overview/Software/Information_Management/Content_Analytics

4.1.3 Determining the installation server type and procedure consideration

Before you install Content Analytics, determine the installation server type. Content Analytics supports the following types of configuration:

- Single server configuration
- Distributed (or multiple) server configuration

This chapter explains a single server installation and configuration. If you plan to perform a distributed server installation, install a master server first and then set up other servers such as a crawler server, document processor server, and search server.

You can add an additional server anytime after a single Content Analytics server or a master server of the distributed Content Analytics server is ready.

System scaling consideration: When you consider scaling your system, add the additional node from the administration console after the additional server is ready. You can configure this additional server as a document processor server, search server, or both.

For additional details, go to the IBM Content Analytics Information Center at the following address, and search on *gathering information for installation*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

4.1.4 Parameters used during the installation

Although the Content Analytics installation program is intuitive and easy to run, you must still consider and determine the parameters to use during the installation.

You might want to change the following typical parameters from the Advanced Options page:

- Installation directory and data directory
- Port assignment
- Administrative user ID and password
- Application server environment

For more details, go to the IBM Content Analytics Information Center at the following address, and search on *gathering information for installation*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

Installation directory and data directory

When you install Content Analytics, you must consider where you want to locate the product-related binary files and collection-related data. Both are referred to as the *installation directory* and *data directory*.

Installation directory

When you install Content Analytics, the installation directory is set by default if you install it in English:

- For AIX, it is in the `/opt/IBM/es` directory.
- For Linux, it is in the `/opt/IBM/es` directory.
- For Windows, it is in the `C:\Program Files\IBM\es` directory.

The installation directory contains important binary or configuration files that never change when the system is running. During the installation, the `ES_INSTALL_ROOT` environment variable is created, and the path that you specified is set. The installation directory path is referred to as `ES_INSTALL_ROOT` throughout this chapter.

Data directory

You must consider where to place the collection-related data or configuration files that will change while the system is running.

When you install Content Analytics, the data directory is set by default:

- For AIX, it is in the `/home/<`*`administrator user name`*`>/esdata` directory.
- For Linux, it is in the `/home/<`*`administrator user name`*`>/esdata` directory.

- For Windows, it is in the `C:\Program Files\IBM\es\<administrator user name>` directory.

For example, AIX uses the `/home/IBM/es/esadmin/esdata` data path when you set the administrator ID as `esadmin`. Similarly, Windows uses the `C:\Program Files/IBM/es/esadmin` data path when you set the administrator ID as `esadmin`.

During the installation, the `ES_NODE_ROOT` environment variable is created, and the path that you specified is set. The data directory path is referred to as `ES_NODE_ROOT` throughout this chapter.

Installation note: In this chapter, the installation is performed with the administrative user on the Windows platform. When you perform a nonroot and nonadministrative user installation on the supported platform, the installation directory path and the data directory path can be different.

You can change the installation directory and data directory during the installation on the Advanced option page. However, make sure that the installation path or the data path does not contain the trailing path delimiter when you do not install the product with the default path. The trailing path delimiter is a forward slash (`/`) for AIX and Linux and a backslash (`\`) for Windows, for example `/usr/IBM/es/` or `C:\Program files\IBM\es\esadmin\`.

For further details, go to the IBM Content Analytics Information Center at the following address, and search on *installation and data directories*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Installation directories for a distributed server installation: Install Content Analytics in the same installation directory when you perform the distributed server installation. You must confirm the installation before you start the master server installation.

Port assignment

Content Analytics uses ports during its processing. Make sure to assign the ports that are not used by the other resources. If the default ports conflict with the other resources, consider changing the ports for Content Analytics, or change the port used by the other resources.

For the default port assignment, go to the IBM Content Analytics Information Center at the following address, and search on *default port assignments*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Administrative user ID and password

During the installation, you have to specify the administrative user ID and password. You can create a new administrative user for Content Analytics during the installation if you start the installation program as the root user or administrative user. Alternatively, you can specify the existing user as the administrative user for Content Analytics.

Throughout this chapter, `esadmin` is used as the administrative user that is the local user ID. That is, the administrative user that belongs to the Windows domain is not used.

> **User ID and distributed server installation:** When you perform the distributed server installation, install the product with the root (or administrative) user and set the same user ID as the administrative user ID for Content Analytics.

You can also install the product as a nonroot user. However, when Content Analytics is installed with a nonroot user, product limitations occur depending on your business requirement. Therefore, it is important to confirm the description in the IBM Content Analytics Information Center. For more information, go to the IBM Content Analytics Information Center at the following address, and search on *administrator ID and password*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Application server environment

Content Analytics installs the Jetty application server by default. The use of IBM WebSphere Application Server is also supported.

If you use WebSphere Application Server instead of the Jetty application server, install the WebSphere Application Server and configure the environment beforehand.

> **Application server:** Throughout this chapter, the Jetty application server is used as the application server that is installed by default.

4.1.5 Installing the agent server

When you want to crawl the files on the remote Windows file server even though you install the Content Analytics server on AIX or Linux, you can install the agent server on a Windows server to crawl the files on the remote Windows file server.

For further details, go to the IBM Content Analytics Information Center at the following address, and search on *installing an agent server*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

4.1.6 Installing Content Analytics on a single server

When you are ready for the installation, you must first execute the installer. This section explains how to install Content Analytics on a Windows 2003 server as a single server configuration with a graphical user interface (GUI) installer.

For further details about the upgrade installation, go to the IBM Content Analytics Information Center at the following address, and search on *upgrading to IBM Content Analytics Version 2.2*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

> **High availability installation:** You can add high availability server after you install a master server or single server. However, high availability configuration is available for AIX and Windows platform only. For further details, go to the IBM Content Analytics Information Center at the following address, and search on *installing additional servers*:
>
> http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Starting the GUI installer

To install Content Analytics with the GUI installer, perform the following steps:

1. Extract the installation image file. You can find the launchpad command in the extracted files. For AIX and Linux, use `launchpad.sh`, and for Windows, use `launchpad.exe`.
2. Run the **`launchpad`** command to start the installation.
3. Select the language to use during the installation (Figure 4-1). For our test scenario, we select **English**. Then click **OK**.

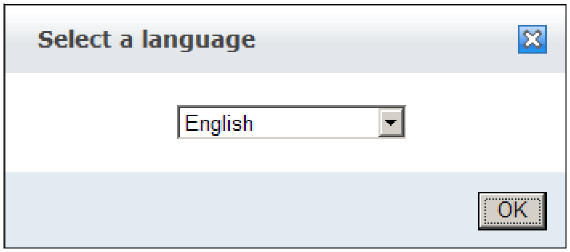
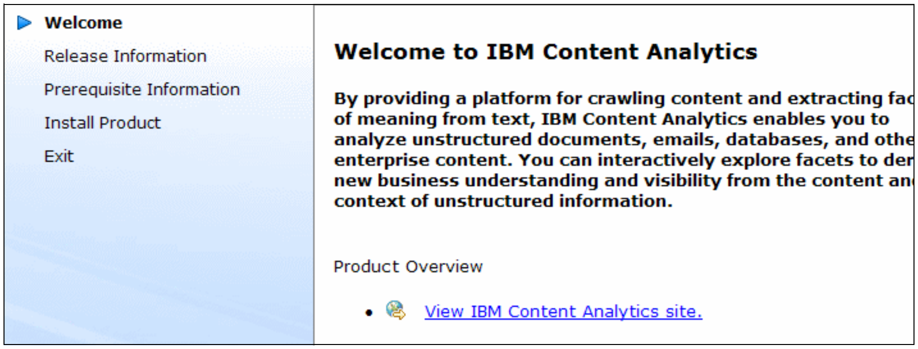


Figure 4-1 Selecting a language that is used during the installation

4. In the Welcome window (Figure 4-2) that opens, in the left pane, click **Install Product**.



Welcome to IBM Content Analytics

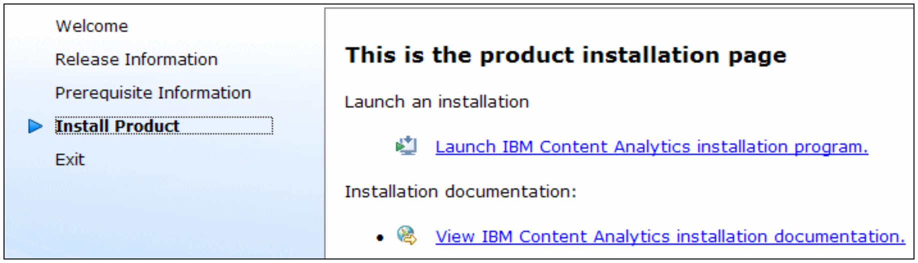
By providing a platform for crawling content and extracting fac of meaning from text, IBM Content Analytics enables you to analyze unstructured documents, emails, databases, and othe enterprise content. You can interactively explore facets to der new business understanding and visibility from the content an context of unstructured information.

Product Overview

- View IBM Content Analytics site.

Figure 4-2   Content Analytics installation: Welcome window

5. In the Install Product window (Figure 4-3) that shows the links to launch the installer, click the **Launch Content Analytics installation program** link.



This is the product installation page

Launch an installation

Launch IBM Content Analytics installation program.

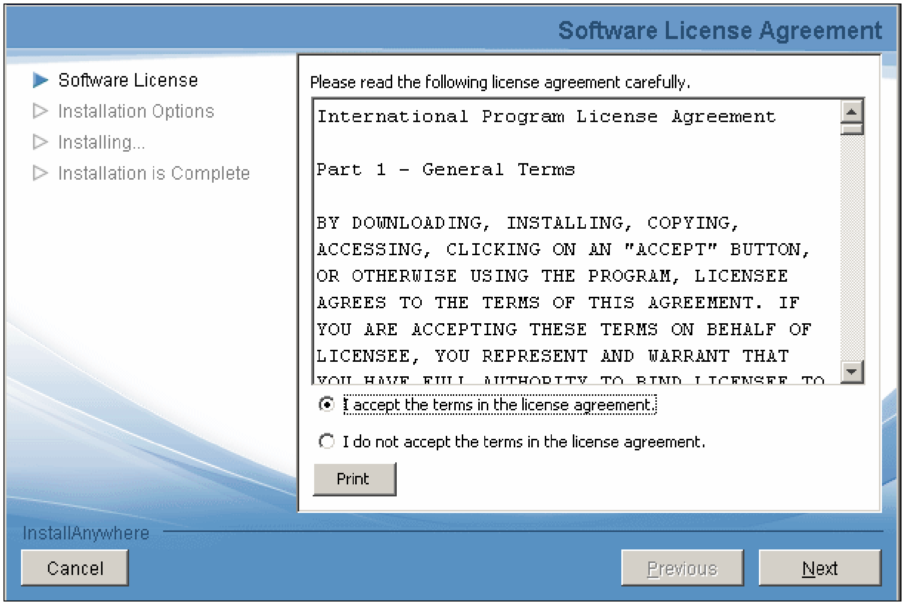
Installation documentation:

- View IBM Content Analytics installation documentation.

Figure 4-3   Content Analytics installation: Install Product window

6. In the window that opens with a banner, click **OK** to proceed.

7. In the Software License window (Figure 4-4), select **I accept the terms in the license agreement** and click **Next**.



Please read the following license agreement carefully.

```
International Program License Agreement

Part 1 - General Terms

BY DOWNLOADING, INSTALLING, COPYING,
ACCESSING, CLICKING ON AN "ACCEPT" BUTTON,
OR OTHERWISE USING THE PROGRAM, LICENSEE
AGREES TO THE TERMS OF THIS AGREEMENT. IF
YOU ARE ACCEPTING THESE TERMS ON BEHALF OF
LICENSEE, YOU REPRESENT AND WARRANT THAT
YOU HAVE FULL AUTHORITY TO BIND LICENSEE TO
```

● I accept the terms in the license agreement.

○ I do not accept the terms in the license agreement.

Figure 4-4 Content Analytics installation: Software License window

8. In the Installation Options window (Figure 4-5 on page 80), specify the following items that you prepared in 4.1.4, “Parameters used during the installation” on page 74:

   – Fully qualified host name

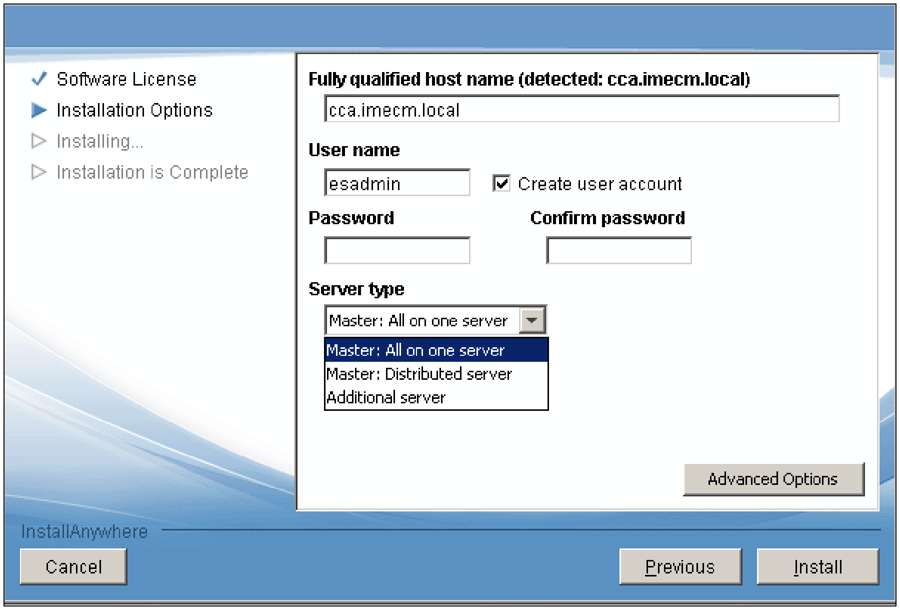
   > **Resolving the host name with a name resolution service:** When you enter the host name, make sure that the host name can be resolved by a name resolution service, such as a domain name server (DNS), or hosts file. Do not specify the IP address as the fully qualified host name.

   – User name and password

   You can create a user or use an existing user for the administrator user.

   – Server type

   If you plan to perform a distributed server installation or additional server installation, select the appropriate server type. For our test scenario, we select **Master: All on one server**.



Fully qualified host name (detected: cca.imecm.local)

cca.imecm.local

User name

esadmin ☑ Create user account

Password

Confirm password

Server type

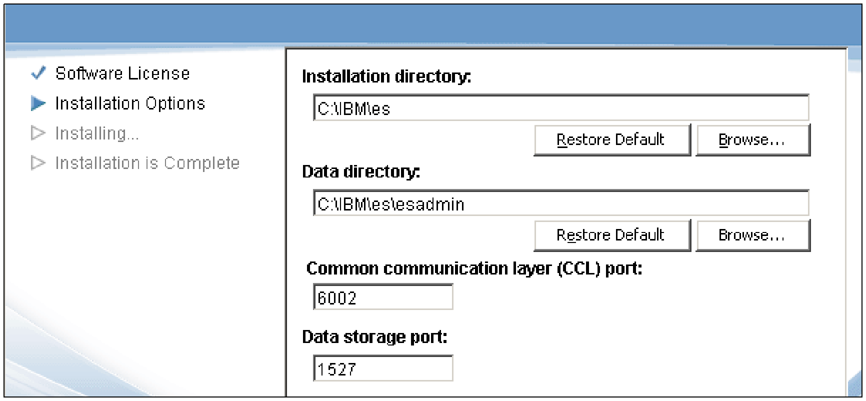
Master: All on one server

Figure 4-5 Content Analytics installation: Installation Options window

If you install the product with the default configuration, click **Install** to start the installation program.

However, if you want to change some of the parameters that were explained in 4.1.4, “Parameters used during the installation” on page 74, select **Advanced Options** in the Installation Options window (Figure 4-5) to change the following parameters:

– The installation and data directories (Figure 4-6)



Installation directory:

C:\IBM\es

Data directory:

C:\IBM\es\esadmin

Common communication layer (CCL) port:

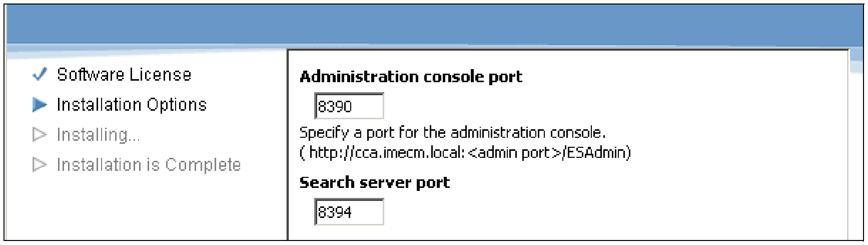
6002

Data storage port:

1527

Figure 4-6 Selecting the installation directory and data directory

– The port numbers for the administration console and search server (Figure 4-7)



Administration console port

8390

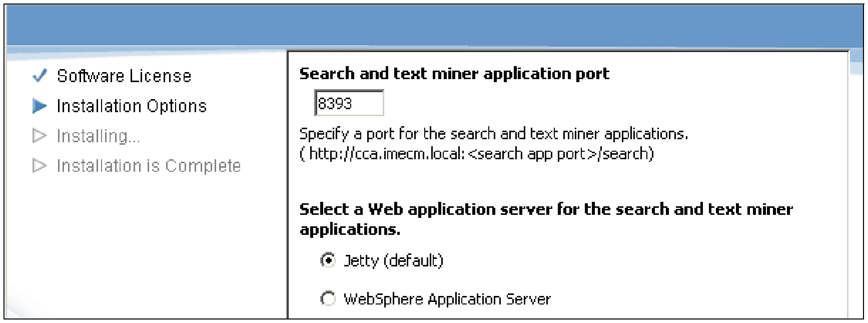
Specify a port for the administration console.
(http://cca.imecm.local: <admin port>/ESAdmin)

Search server port

8394

Figure 4-7 Specifying the administration console and search server port

– The port numbers for the search and text miner application and selection of the web application server (Figure 4-8)



Search and text miner application port

8393

Specify a port for the search and text miner applications.
(http://cca.imecm.local: <search app port>/search)

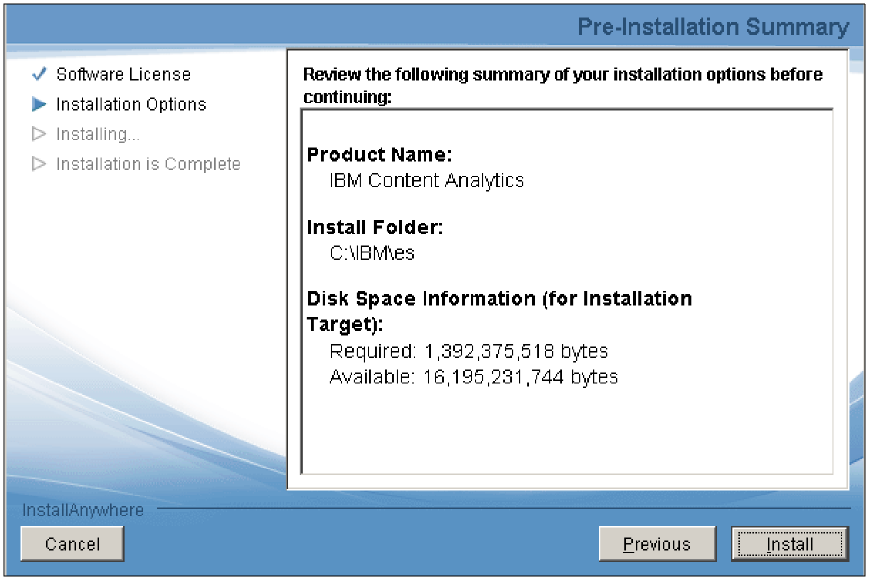
Select a Web application server for the search and text miner applications.

● Jetty (default)

○ WebSphere Application Server

Figure 4-8 Specifying the text miner application port and application server

When you select all the parameters, the summary window (Figure 4-9) is displayed. Confirm your settings and the amount of disk space that is used for the installation. Then click **Install** to start the installation program.



Review the following summary of your installation options before continuing:

Product Name:
IBM Content Analytics

Install Folder:
C:\IBM\es

Disk Space Information (for Installation Target):
Required: 1,392,375,518 bytes
Available: 16,195,231,744 bytes

Figure 4-9 Content Analytics installation: Pre-Installation Summary window

For more details, go to the IBM Content Analytics Information Center at the following address, search on *performing a fresh installation of IBM Content Analytics*, and review its subtopics:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

4.1.7 Running the First Steps program to verify the installation

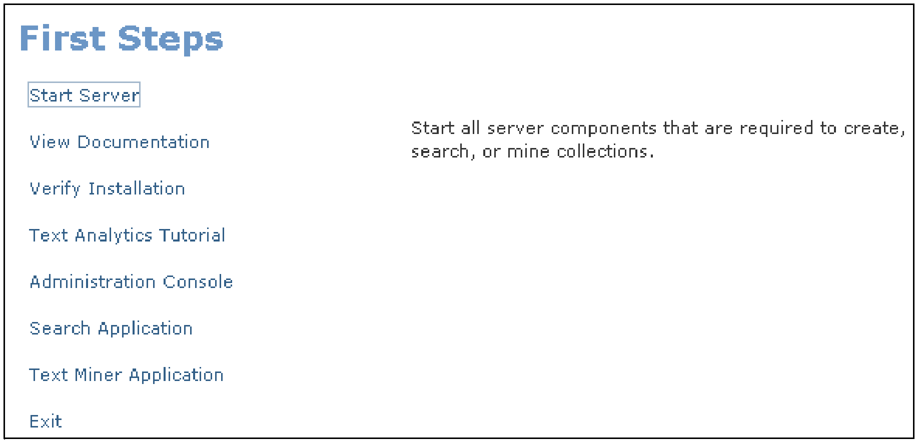
The First Steps program helps you to verify whether the installation is successful. After you finish the installation, run the First Steps program.

Installing Content Analytics on Windows 2008 server: To start Content Analytics, you must log in as the Content Analytics administrative user (`esadmin` in the previous example) at least once after the installation is complete. If you log in as the Windows 2008 system administrator and do not log in as the Content Analytics administrative user after the installation, errors might occur while starting the server.

To start the first Steps program, perform the following steps:

1. Choose one of the following methods depending on your environment:
   - For the AIX and Linux platform, run **`firststep.sh`**.
   - For the Windows platform, run **`firststep.bat`** from the command line.

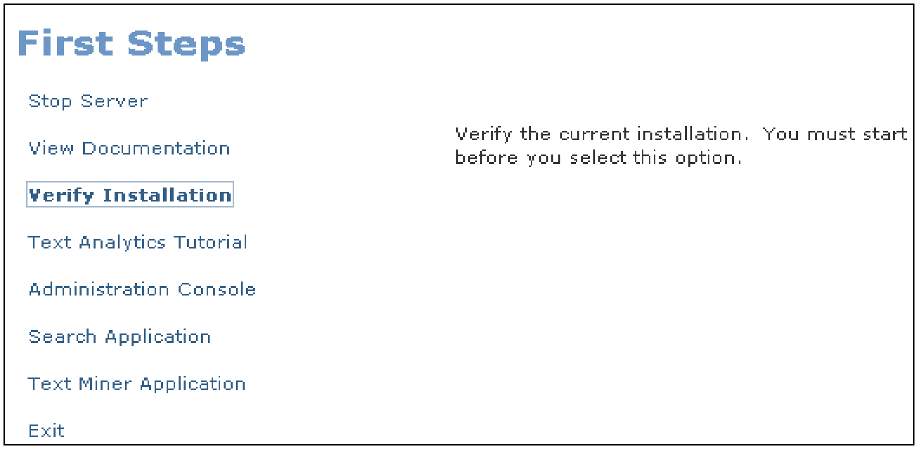
     If you installed Content Analytics on a Windows platform, select **Start** → **All Programs** → **IBM Content Analytics** → **First Steps**.
2. In the Welcome window (Figure 4-10), click **Start Server**.



Start all server components that are required to create, search, or mine collections.

Figure 4-10 First Steps: Choosing the Start Server option

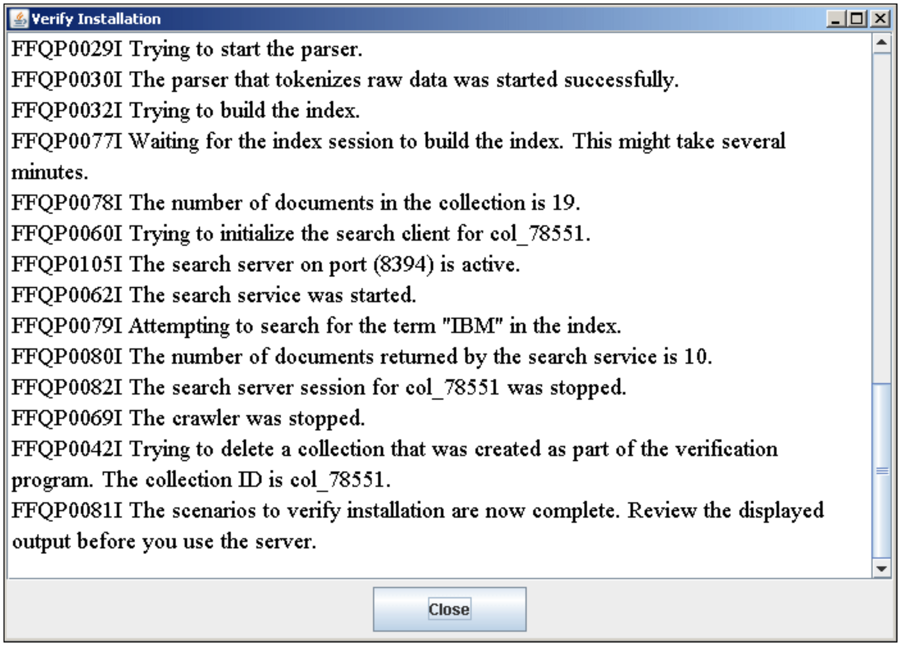
3. After starting the server, click **Verify Installation** to verify the current installation (Figure 4-11).



Verify the current installation. You must start before you select this option.

Figure 4-11 First Steps: Choosing the Verify Installation option

The verification program opens the Verify Installation result window (Figure 4-12). Any problems are reported in this window. If you do not see any error messages, the installation is successful.



```
FFQP0029I Trying to start the parser.
FFQP0030I The parser that tokenizes raw data was started successfully.
FFQP0032I Trying to build the index.
FFQP0077I Waiting for the index session to build the index. This might take several
minutes.
FFQP0078I The number of documents in the collection is 19.
FFQP0060I Trying to initialize the search client for col_78551.
FFQP0105I The search server on port (8394) is active.
FFQP0062I The search service was started.
FFQP0079I Attempting to search for the term "IBM" in the index.
FFQP0080I The number of documents returned by the search service is 10.
FFQP0082I The search server session for col_78551 was stopped.
FFQP0069I The crawler was stopped.
FFQP0042I Trying to delete a collection that was created as part of the verification
program. The collection ID is col_78551.
FFQP0081I The scenarios to verify installation are now complete. Review the displayed
output before you use the server.
```

Figure 4-12 First Steps: Verify Installation results window

4.1.8 Starting the Text Analytics tutorial

From the First Steps program, you can start the Text Analytics tutorial. When you select the Text Analytics tutorial, the Sample Text Analytics Collection is created. In 4.3.2, “Creating a text analytics collection” on page 90, you see how to create the text analytics collection by yourself. If you want to see the text miner application, you can create the Sample Text Analytics Collection from here.

4.2 Administering Content Analytics

After you successfully install Content Analytics, you create a text analytics collection to store the data to analyze. This section addresses the basic administration operations such as how to start Content Analytics, how to access the administration console, and how to stop Content Analytics.

4.2.1 Starting the system

Before you configure Content Analytics, you must start Content Analytics. To start Content Analytics, you log in as an administrative user and issue the command to start the entire system.

To start Content Analytics, follow these steps:

1. Start the common communication layer (CCL) by entering the following command:

   ```
   startccl
   ```

   Note the following considerations:

   – If you use a single server configuration, you do not need this step because the **`esadmin system startall`** command starts the CCL.

   – If you use a distributed server configuration, start the CCL on all servers beforehand.

2. Start the system components by entering the following command:

   ```
   esadmin system startall
   ```

   Alternatively, if you install Content Analytics on a Windows platform, select **Start** → **All Programs** → **IBM Content Analytics** → **Start Up**. The command starts the CCL sessions and other services such as the text miner application.

Starting the CCL:

- If you install Content Analytics on the Windows platform, you can start the CCL automatically as a Windows service. Because the entire Content Analytics system does not start when you start CCL, you must still start the entire system by using the **`esadmin system startall`** command.
- If you have a distributed server configuration environment, enter the **`esadmin system startall`** command from the master server after CCL is started on all nodes.

4.2.2 Accessing the administration console

To create and configure a text analytics collection, you must access the administration console from following URL:

```
http://<Content Analytics server host name:Port Number>/ESAdmin/
```

The *`host name`* is the server on which you install Content Analytics. The *`port number`* is the number that you specified during installation. By default, the port number is 8390. If you install Content Analytics as a distributed server installation, you must specify the Content Analytics server host name as the master server.

When the login page opens, specify the administrative user name and password that is specified during the installation.

4.2.3 Stopping the server

Sometimes you might need to stop Content Analytics. To stop Content Analytics, log in as an administrative user and enter the following command:

```
esadmin system stopall
```

This command stops all Content Analytics-related components. Alternatively, if you install Content Analytics on a Windows platform, select **Start** → **All Programs** → **IBM Content Analytics** → **Shut Down**.

Although the multiserver configuration is not explained in this chapter, you might want to start or stop a specific server. To start a specific server, you use the **`esadmin system startLocal`** command. To stop a specific server, you use the **`esadmin system stopLocal`** command.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *starting and stopping local services*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

4.3 Configuring a text analytics collection

Now that you have installed Content Analytics and verified that it is accessible, the next step is to build a text analytics collection. In this section, you learn how to design and create the Sample Text Analytics Collection with sample data provided by Content Analytics. The instructions are based on a single server configuration (master, all-on-one server setup) in the Windows environment created in 4.2, “Administering Content Analytics” on page 84. You use the sample data in the `*ES_INSTALL_ROOT*/samples/firststep` directory.

To build a text analytics collection, you perform the following tasks as explained in the sections that follow:

1. Designing a sample collection
2. Creating a text analytics collection
3. Defining and configuring a crawler
4. Building an index in the text analytics collection

Text Analytics Tutorial: If you do not want to walk through the steps for building a text analytics collection process, click **Text Analytics Tutorial** on the First Steps pane to automatically build a collection. The installation and data directories defined during the installation (Figure 4-6 on page 80) are automatically used, and a new collection called Sample Text Analytics Collection is created. For instructions on accessing the First Steps program, see the 4.1.7, “Running the First Steps program to verify the installation” on page 82.

4.3.1 Designing a sample collection

Planning and designing a text analytics collection is one of the most important steps and requires several iterations to achieve the desired result. Chapter 2, “Application design and preparation” on page 27, explains how to plan and design an application. This section applies the definitions, concepts, and tasks defined in that chapter on the sample data provided by Content Analytics to plan, design, and then build a sample collection.

Designing a collection entails performing the following tasks as explained in this section:

1. Examining input data and identifying the native fields
2. Identifying search fields from native fields
3. Mapping native fields to search fields
4. Identifying facets and mapping search fields to facets

Examining input data and identifying the native fields

You use the sample data that is provided by Content Analytics in the test scenario. This data represents the inbound call center data of a fictitious confectionery Company A. The archive file is in the `ES_INSTALL_ROOT\samples\firststep\data\xml\xmls.tar.gz` directory.

Extract the compressed file anywhere on your machine. For this testing scenario, the data is extracted into the `ES_INSTALL_ROOT\samples\firststep\data\xml\xml-data` directory.

Example 4-1 shows one of the documents from the sample data set.

Example 4-1 Format of the sample data

```
<?xml version="1.0" encoding="UTF-8"?>
<doc>
    <id>00000000</id>
    <title>lemon tea - Package / container</title>
    <date>2008-01-01</date>
    <timestamp>1199186392296</timestamp>
    <category>Package / container</category>
    <subcategory>Straw</subcategory>
    <product>lemon tea</product>
    <text>[Pack] The straw was peeled off from the juice pack.</text>
</doc>
```

The native fields are the metadata or properties that describe a document. In the sample data set, the native fields (id, title, date, timestamp, category, subcategory, and product) are identified. The <text> field contains the unstructured data that is considered content or textual data of the document.

Identifying search fields from native fields

Search fields are generally a common attribute among data sets. For example, date, title, subject, and owner are some of the common properties shared by a majority of the data sets. By default, Content Analytics defines these search fields. However, in most cases, the user must define more search fields based on the data set. For each native field that you identify, begin by asking questions such as: What are you interested in from the set of data? Do you want to search by product? Do you want to narrow down the result by category? If the answer to these questions is “yes,” ensure that the native field is mapped to a search field.

In our testing scenario, we select the following common fields of interest:

- category
- id
- product
- subcategory
- timestamp
- title
- date

After you identify the fields that you want to map as search fields, see if Content Analytics provides any default search fields that can map to these fields. Content Analytics provides the following search fields by default:

- author
- body
- createddate
- database
- databasepath
- databasetitle
- date
- directory

- extension
- filename
- filesize
- keywords
- modifieddate
- page
- table
- title
- typename
- version
- view

Based on this list, you can map the native fields of title and date to the default title and date search fields provided by Content Analytics. You need to create new search fields for the other native fields, including id, category, subcategory, timestamp, and product. In “Creating the search fields” on page 96, you create the following search fields:

- doc_category
- doc_date
- doc_id
- doc_product
- doc_subcategory

Mapping native fields to search fields

Based on the native fields and search fields that you identify, you can create a mapping of XML metadata (native fields) to the search fields as shown in Table 4-1.

Table 4-1 Mapping the native fields to the search fields

Native field from an original document (XML metadata)	Search field in Content Analytics
id	doc_id[a]
title	title
date	date
category	doc_category[a]
subcategory	doc_subcategory[a]
product	doc_subproduct[a]
timestamp	doc_date

[a] You must create this search field before you do the mapping.

Identifying facets and mapping search fields to facets

Based on the native fields that you identify, we want to narrow down the search results by a certain category, subcategory, or product. Thus, you create facets and map them to the search fields as listed in Table 4-2. Additionally, you create a facet for the doc_date field to illustrate how you can create multiple ranges in a facet for fields that are of the date (or decimal) type.

Table 4-2 Mapping the search field to the facet

Facet	Search field
Category	doc_category
Subcategory	doc_subcategory
Product	doc_product
Report Date	doc_date

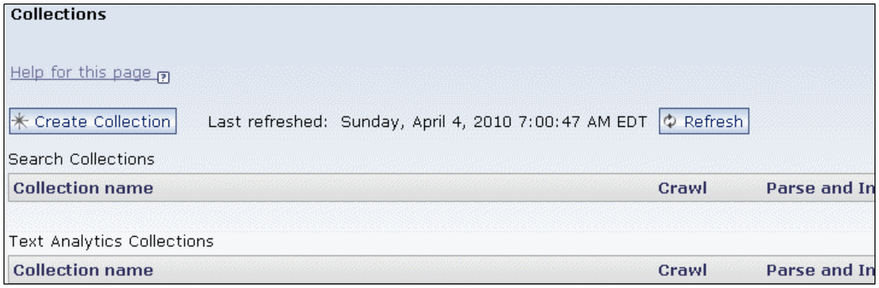
4.3.2 Creating a text analytics collection

Creating a text analytics collection entails the following tasks as explained in this section:

1. Creating the collection
2. Creating the search fields
3. Mapping the native field to the search field
4. Creating facets and mapping search fields to facets
5. Optional: Configuring the date facet

Creating the collection

The first step in analyzing data is to create a text analytics collection by using the administration console. After you log in to the administration console, the main collection table is displayed in the Collections window (Figure 4-13). When you log in the first time, no collection is created.



Collections

Help for this page

Last refreshed: Sunday, April 4, 2010 7:00:47 AM EDT

Search Collections

Collection name	Crawl	Parse and In

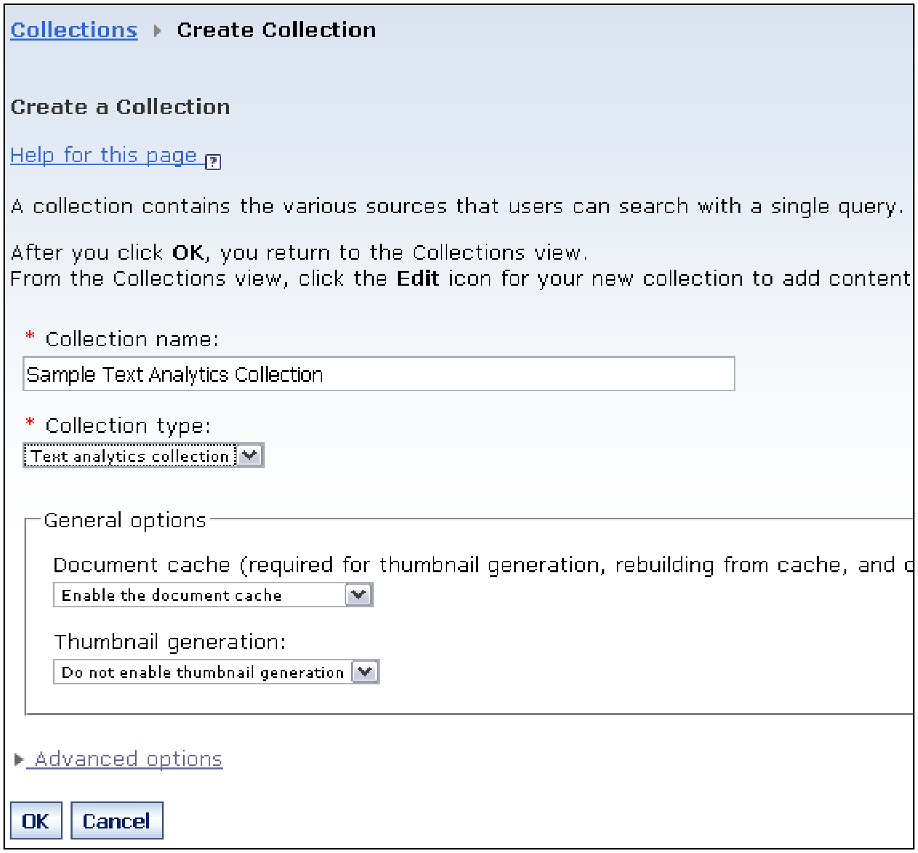
Text Analytics Collections

Collection name	Crawl	Parse and In

Figure 4-13 Collections view window

To create a text analytics collection, follow these steps:

1. In the Collection window (Figure 4-13 on page 90), click **Create Collection**.
2. In the Create Collection window (Figure 4-14), specify the following values:
   a. For Collection name, enter `Sample Text Analytics Collection`.
   b. For Collection type, select **Text analytics collection**.



Collections ▸ Create Collection

Create a Collection

Help for this page

A collection contains the various sources that users can search with a single query.

After you click **OK**, you return to the Collections view.
From the Collections view, click the **Edit** icon for your new collection to add content

* Collection name:
Sample Text Analytics Collection

* Collection type:
Text analytics collection

General options

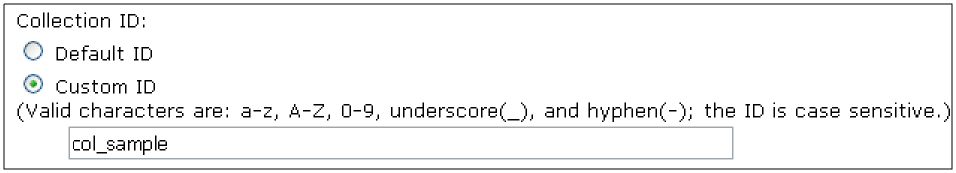
Document cache (required for thumbnail generation, rebuilding from cache, and c
Enable the document cache

Thumbnail generation:
Do not enable thumbnail generation

Figure 4-14 Create Collection window

c. Click **Advanced options** and set the following information:

i. For Collection ID field, select **Custom ID** and enter `col_sample` (Figure 4-15). Alternatively, you can leave the setting as **Default ID**, and the system assigns an ID automatically.



Collection ID:

○ Default ID

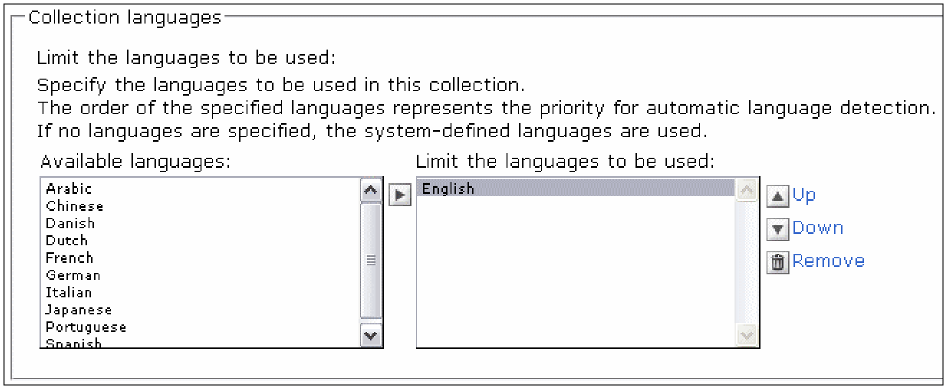
● Custom ID

(Valid characters are: a-z, A-Z, 0-9, underscore(_), and hyphen(-); the ID is case sensitive.)

col_sample

Figure 4-15 Assign Collection ID window

ii. For Collection languages (Figure 4-16), select **English** from the left column and click the **arrow** icon to move the selected language to the right column.



Collection languages

Limit the languages to be used:

Specify the languages to be used in this collection.
The order of the specified languages represents the priority for automatic language detection.
If no languages are specified, the system-defined languages are used.

Available languages:

Arabic
Chinese
Danish
Dutch
French
German
Italian
Japanese
Portuguese
Spanish

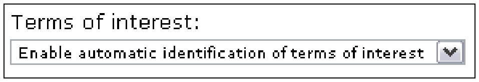
Limit the languages to be used:

English

Figure 4-16 Collection Languages window

iii. Enable Terms of interest by selecting **Enable automatic identification of terms of interest** (Figure 4-17), which is a new feature in Content Analytics Version 2.2. This option identifies entities and predicates that might be of interest based on analyzing textual content of the collection.

> **More information:** For further details about the terms of interest functionality, see 8.1, “The power of dictionary-driven analytics” on page 322, and 8.2, “Terms of interest” on page 326.



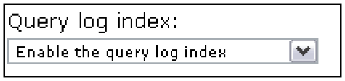
Terms of interest:

Enable automatic identification of terms of interest

Figure 4-17 Enabling the terms of interest feature for the collection

iv. Ensure that **Enable query log index** is selected for the Query log index field (Figure 4-18). The query log index contains a list of previously executed queries. The search server uses these queries to make suggestions as part of the type-ahead feature.

> **Type-ahead feature:** For further details on the type-ahead functionality, see 5.2.5, “Type ahead” on page 157.

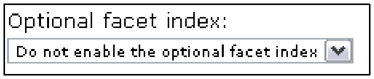


Query log index:

Enable the query log index

Figure 4-18 Enabling the query log index for the collection

v. Enable or disable the optional facet index feature when you create the collection. When you enable this feature, it creates an index to store facets. In this example, we do not create the optional facet index. Therefore, select **Do not enable the optional facet index** as shown in Figure 4-19.



Optional facet index:

Do not enable the optional facet index

Figure 4-19 Enabling the optional facet index for the collection

Building an optional facet index: The optional facet index improves text miner performance (related to response speed) of the text miner application for a large amount of data. However, because a new index is created specifically for the facets, it takes longer to index a document when this feature is enabled. Despite this trade-off, use this option when you have a large collection because it improves text miner performance.

The optional facet index requires you to rebuild the entire index after each data update, and it requires additional disk space. (The optional facet index is a separate index that is built when you follow the steps in 4.3.4, “Building an index in the text analytics collection” on page 132.) The optional facet index is not suitable when your data is frequently updated or when you want to build an index incrementally.

We do not create the optional facet index in this test scenario because our data size is relatively small. You might want to create the optional facet index when your environment meets the situation as described.

For more information about how to build the optional facet index, see 15.5.2, “Enabling the optional facet index” on page 591.

vi. Set the time zone for the collection.

Setting the correct collection time zone (Advanced options): You can select a time zone to represent the date facet in the Advanced options section when creating a collection.

Setting the proper time zone information is important when you analyze data based on a date, such as selecting **Day** as a time scale in the Time Series view. (See Chapter 5, “Text miner application: Basic features” on page 143, for more details.) You might see a difference in the day represented by the date value if the time zone information is not properly configured.

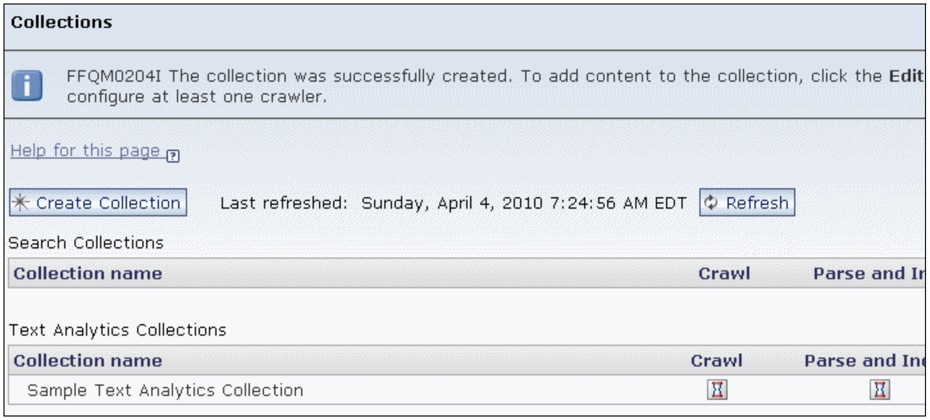
For example, suppose you have three documents with a creation date of 30 April at 1:00 AM according to the time zone in Japan (GMT+9) within the data source. You set the collection to a US Eastern time zone (GMT-5). Content Analytics views these documents with a creation date of April 29 due to the different time zone used by the collection. As a result, the document data might be displayed in the Time Series graph (by creation date) for April 29. Therefore, make sure to select the correct time zone that represents your data in the collection configuration.

vii. Optional: Select additional advanced settings such as choosing the custom location of the data and enabling security. To understand all the advanced options, click **Help for this page**.

Configuring the advanced options: You can enable or disable the optional facet index or terms of interest features after creating the collection from the **General** tab on the administration console. However, you cannot disable the query log index feature after the collection is created.

d. Click **OK**.

Figure 4-20 shows the collection that is created.



Collections

FFQM0204I The collection was successfully created. To add content to the collection, click the **Edit**
configure at least one crawler.

Help for this page

✳ Create Collection   Last refreshed: Sunday, April 4, 2010 7:24:56 AM EDT   Refresh

Search Collections

Collection name	Crawl	Parse and Ir

Text Analytics Collections

Collection name	Crawl	Parse and In
Sample Text Analytics Collection	⌛	⌛

Figure 4-20 Sample Text Analytics Collection

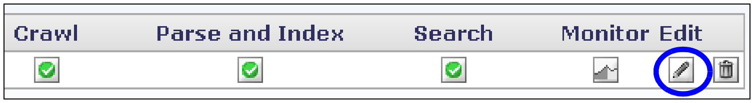
Parse, index, and search components: The parse, index, and search components start automatically. When you create a crawler and start the crawler component, the system automatically parses and indexes the incoming documents. If you do not want this behavior, or if you want to conserve resources on the system, stop the parse, index, and search components.

Creating the search fields

After you create the text analytics collection, you are ready to create the search fields based on the fields that you identified in “Identifying search fields from native fields” on page 88.

To create the search fields, follow these steps:

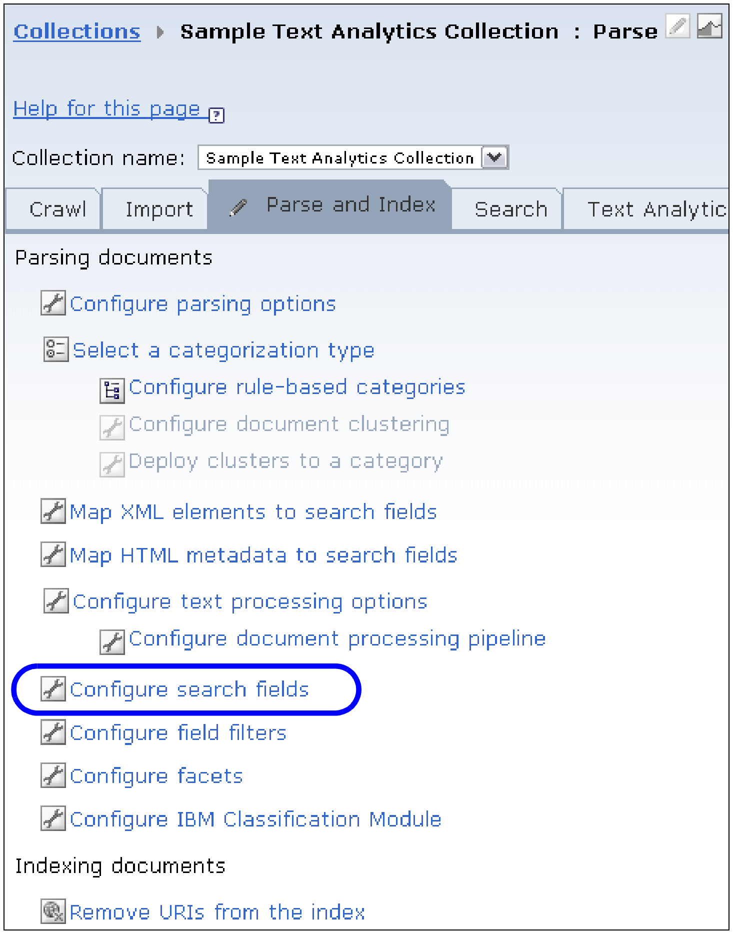
1. Log in to the administration console and click **Collections** in the toolbar to open the Collections view.
2. In the list of collections, locate the collection that you want to edit, and click the **Edit** icon (Figure 4-21).



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓	📈	✎ 🗑

Figure 4-21 Editing and monitoring options for the collection

3. Select the **Parse and Index** tab and then click **Configure search fields** (Figure 4-22).



Collections ▸ Sample Text Analytics Collection : Parse

Help for this page

Collection name: Sample Text Analytics Collection

Parsing documents

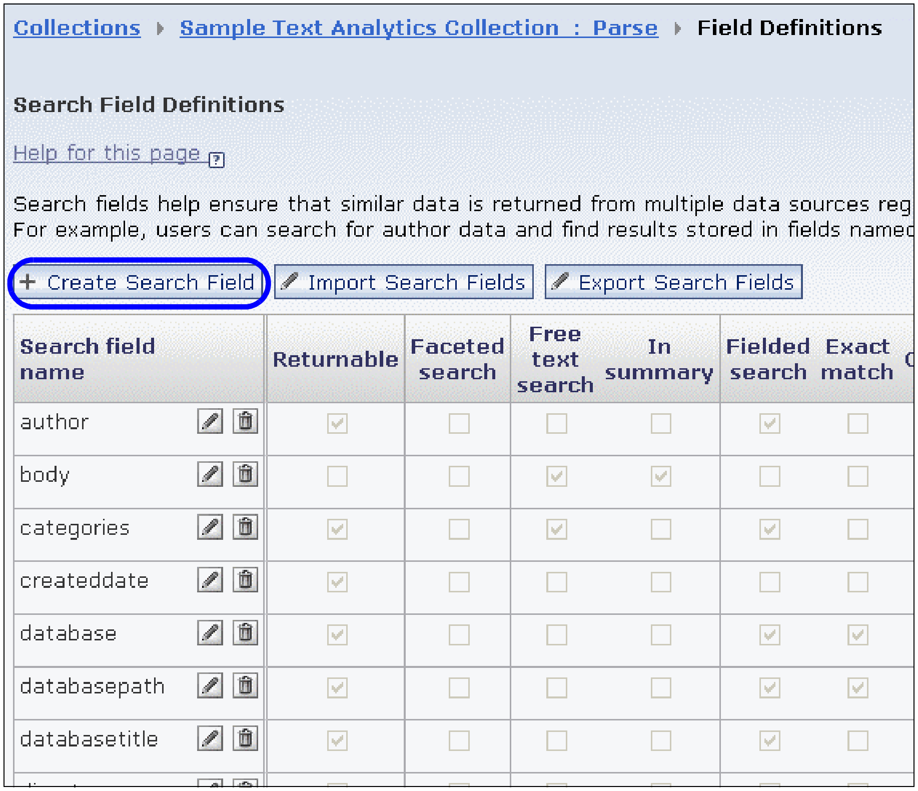
- Configure parsing options
- Select a categorization type
  - Configure rule-based categories
  - Configure document clustering
  - Deploy clusters to a category
- Map XML elements to search fields
- Map HTML metadata to search fields
- Configure text processing options
  - Configure document processing pipeline
- Configure search fields
- Configure field filters
- Configure facets
- Configure IBM Classification Module

Indexing documents

- Remove URIs from the index

Figure 4-22 Parse and Index tab

4. In the Search Fields Definitions window (Figure 4-23), click **Create Search Field**.



Collections ▸ Sample Text Analytics Collection : Parse ▸ Field Definitions

Search Field Definitions

Help for this page

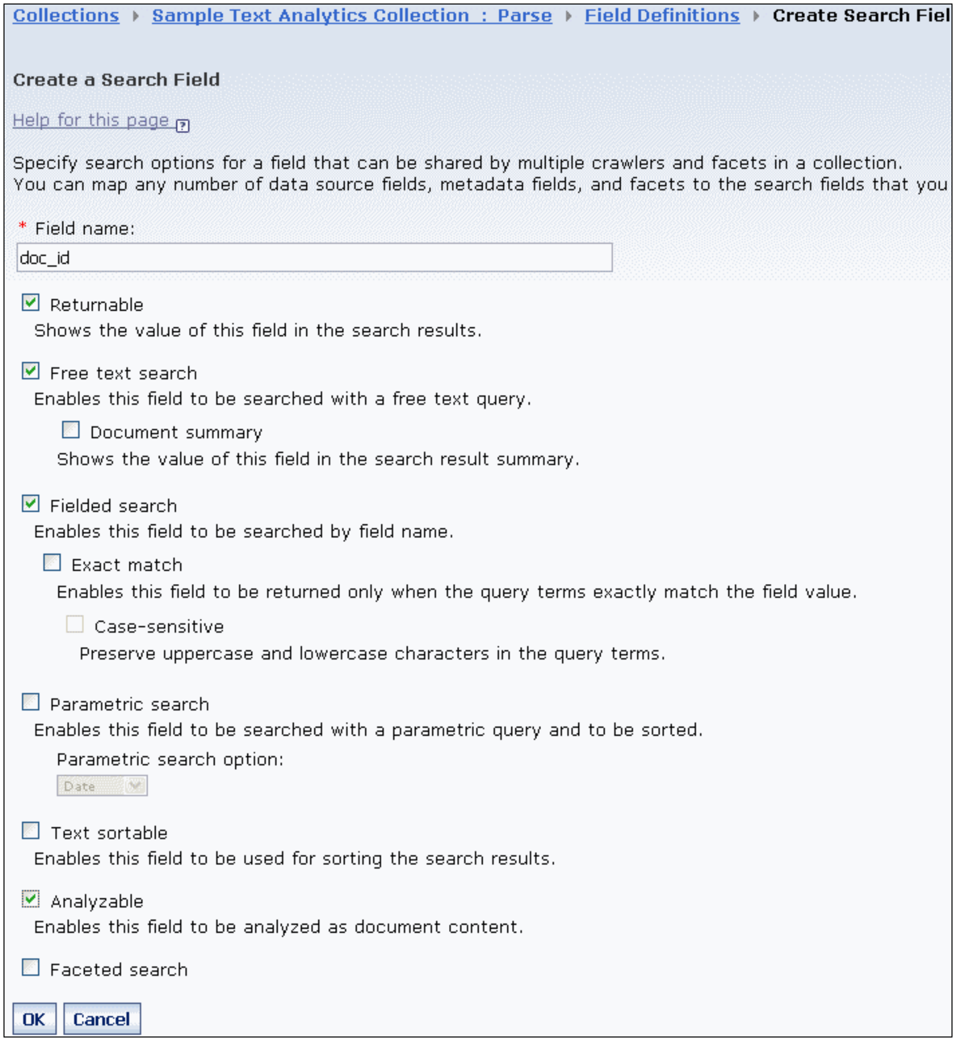
Search fields help ensure that similar data is returned from multiple data sources reg
For example, users can search for author data and find results stored in fields namec

Search field name	Returnable	Faceted search	Free text search	In summary	Fielded search	Exact match
author	☑	□	□	□	☑	□
body	□	□	☑	☑	□	□
categories	☑	□	☑	□	☑	□
createddate	☑	□	□	□	□	□
database	☑	□	□	□	☑	☑
databasepath	☑	□	□	□	☑	☑
databasetitle	☑	□	□	□	☑	□

Figure 4-23 Default Search Field Definitions window

5. In the Create a Search Field window (Figure 4-24 on page 99), for the field name, enter `doc_id`. Select the **Returnable**, **Free text search**, **Fielded search**, and **Analyzable** check boxes, and click **OK**.

> **Faceted search:** Consider these points when creating a search field:
>
> - If you select the **Faceted search** check box when creating a field, a new facet with the same name as the search field is created. However, you cannot change the name of the facet nor group this facet with other facets.
> - After the search field is created, you cannot modify its definition to add facet by selecting the **Faceted search** check box. To map a new facet to the search field, you must manually create the facet and then map it to the existing search field by using the facet panel.



Collections ▸ Sample Text Analytics Collection : Parse ▸ Field Definitions ▸ Create Search Fiel

Create a Search Field

Help for this page

Specify search options for a field that can be shared by multiple crawlers and facets in a collection.
You can map any number of data source fields, metadata fields, and facets to the search fields that you

* Field name:
doc_id

☑ Returnable
Shows the value of this field in the search results.

☑ Free text search
Enables this field to be searched with a free text query.

☐ Document summary
Shows the value of this field in the search result summary.

☑ Fielded search
Enables this field to be searched by field name.

☐ Exact match
Enables this field to be returned only when the query terms exactly match the field value.

☐ Case-sensitive
Preserve uppercase and lowercase characters in the query terms.

☐ Parametric search
Enables this field to be searched with a parametric query and to be sorted.

Parametric search option:
Date

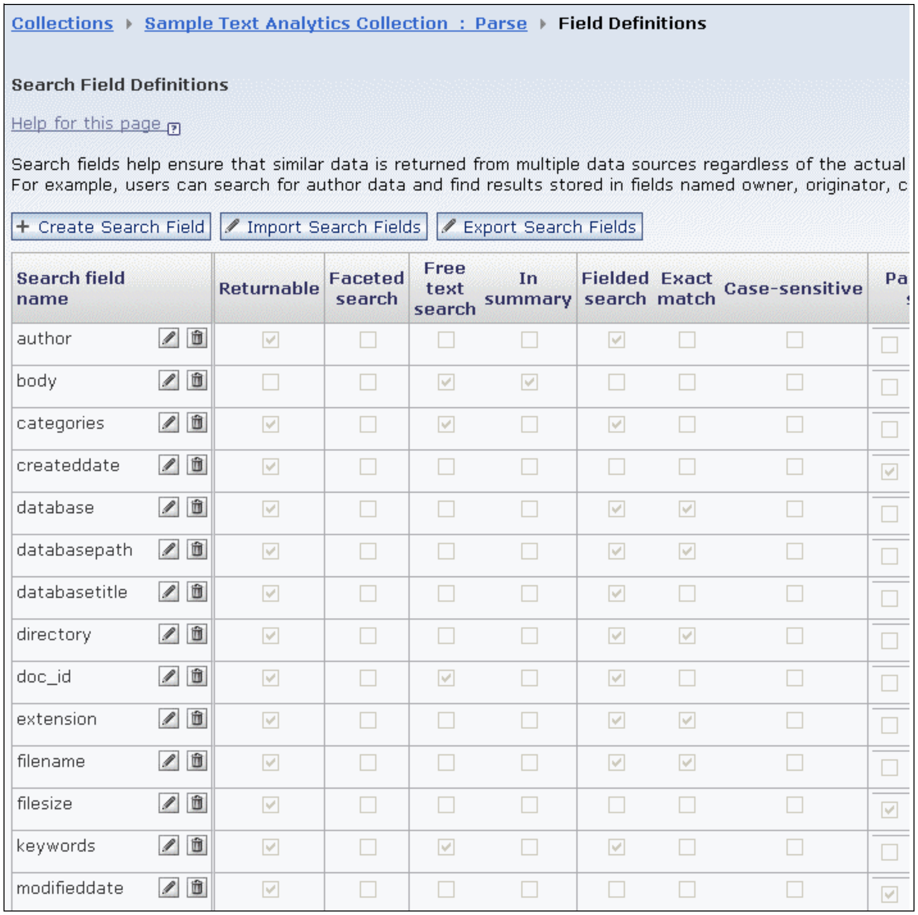
☐ Text sortable
Enables this field to be used for sorting the search results.

☑ Analyzable
Enables this field to be analyzed as document content.

☐ Faceted search

Figure 4-24 Creating a Search Field window

A new search field is created in the Search Field Definitions window (Figure 4-25).



Collections ▸ Sample Text Analytics Collection : Parse ▸ **Field Definitions**

Search Field Definitions

Help for this page

Search fields help ensure that similar data is returned from multiple data sources regardless of the actual
For example, users can search for author data and find results stored in fields named owner, originator, c

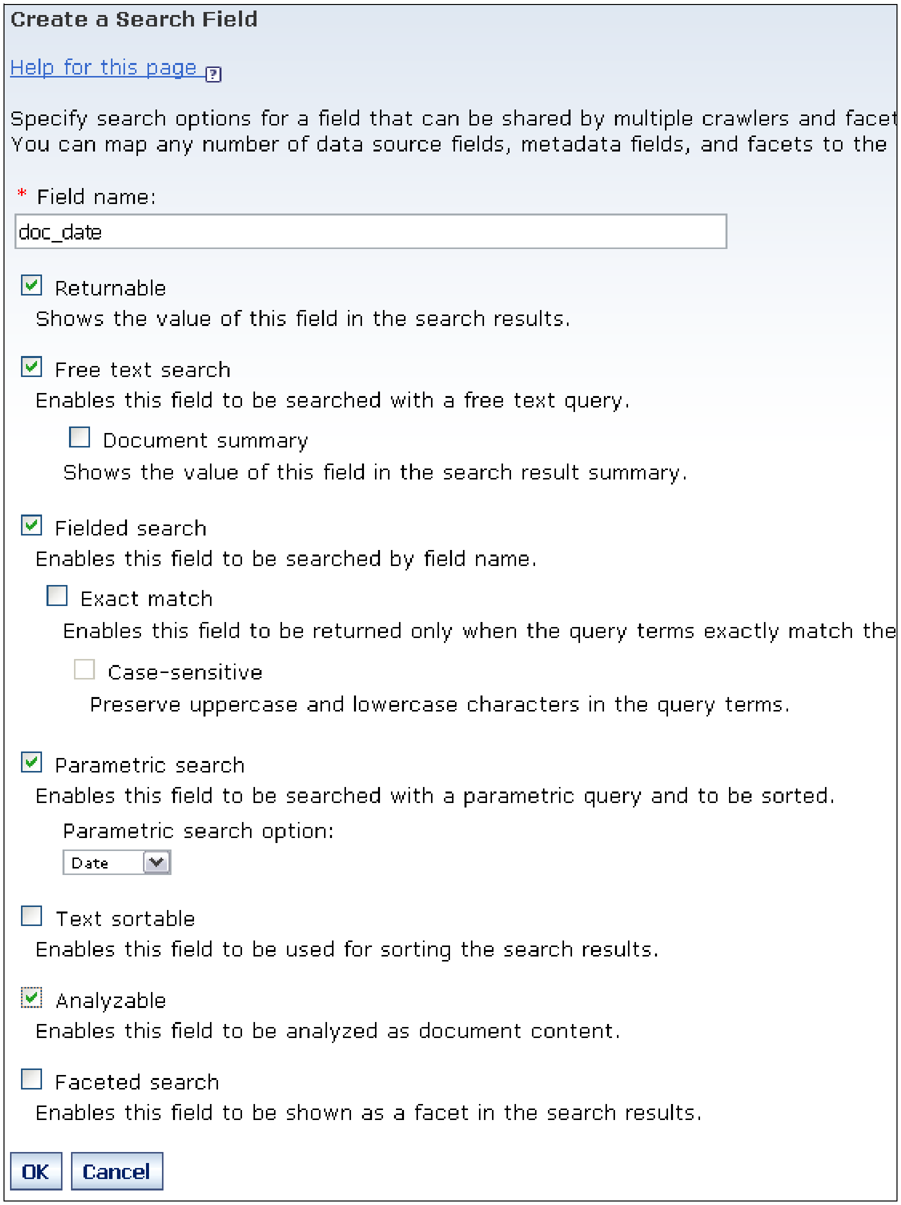
Search field name	Returnable	Faceted search	Free text search	In summary	Fielded search	Exact match	Case-sensitive	Pa
author	☑	☐	☐	☐	☑	☐	☐	☐
body	☐	☐	☑	☑	☐	☐	☐	☐
categories	☑	☐	☑	☐	☑	☐	☐	☐
createddate	☑	☐	☐	☐	☐	☐	☐	☑
database	☑	☐	☐	☐	☑	☑	☐	☐
databasepath	☑	☐	☐	☐	☑	☑	☐	☐
databasetitle	☑	☐	☐	☐	☑	☐	☐	☐
directory	☑	☐	☐	☐	☑	☑	☐	☐
doc_id	☑	☐	☑	☐	☑	☐	☐	☐
extension	☑	☐	☐	☐	☑	☑	☐	☐
filename	☑	☐	☐	☐	☑	☑	☐	☐
filesize	☑	☐	☐	☐	☐	☐	☐	☑
keywords	☑	☐	☑	☐	☑	☐	☐	☐
modifieddate	☑	☐	☐	☐	☐	☐	☐	☑

Figure 4-25 The newly created doc_id search field

6. Repeat step 4 on page 98 through step 5 on page 98 and create the following search fields:

   – doc_category
   – doc_subcategory
   – doc_product

7. Repeat step 4 on page 98 through step 5 on page 98 to create the search field named `doc_date`. However, select the **Parametric Search** check box in addition to selecting the **Returnable**, **Free text search**, **Fielded search**, and **Analyzable** check boxes when creating the field. For Parametric search options, select Date (Figure 4-26).



Create a Search Field

Help for this page

Specify search options for a field that can be shared by multiple crawlers and facet
You can map any number of data source fields, metadata fields, and facets to the

* Field name:
doc_date

☑ Returnable
Shows the value of this field in the search results.

☑ Free text search
Enables this field to be searched with a free text query.

☐ Document summary
Shows the value of this field in the search result summary.

☑ Fielded search
Enables this field to be searched by field name.

☐ Exact match
Enables this field to be returned only when the query terms exactly match the

☐ Case-sensitive
Preserve uppercase and lowercase characters in the query terms.

☑ Parametric search
Enables this field to be searched with a parametric query and to be sorted.

Parametric search option:
Date

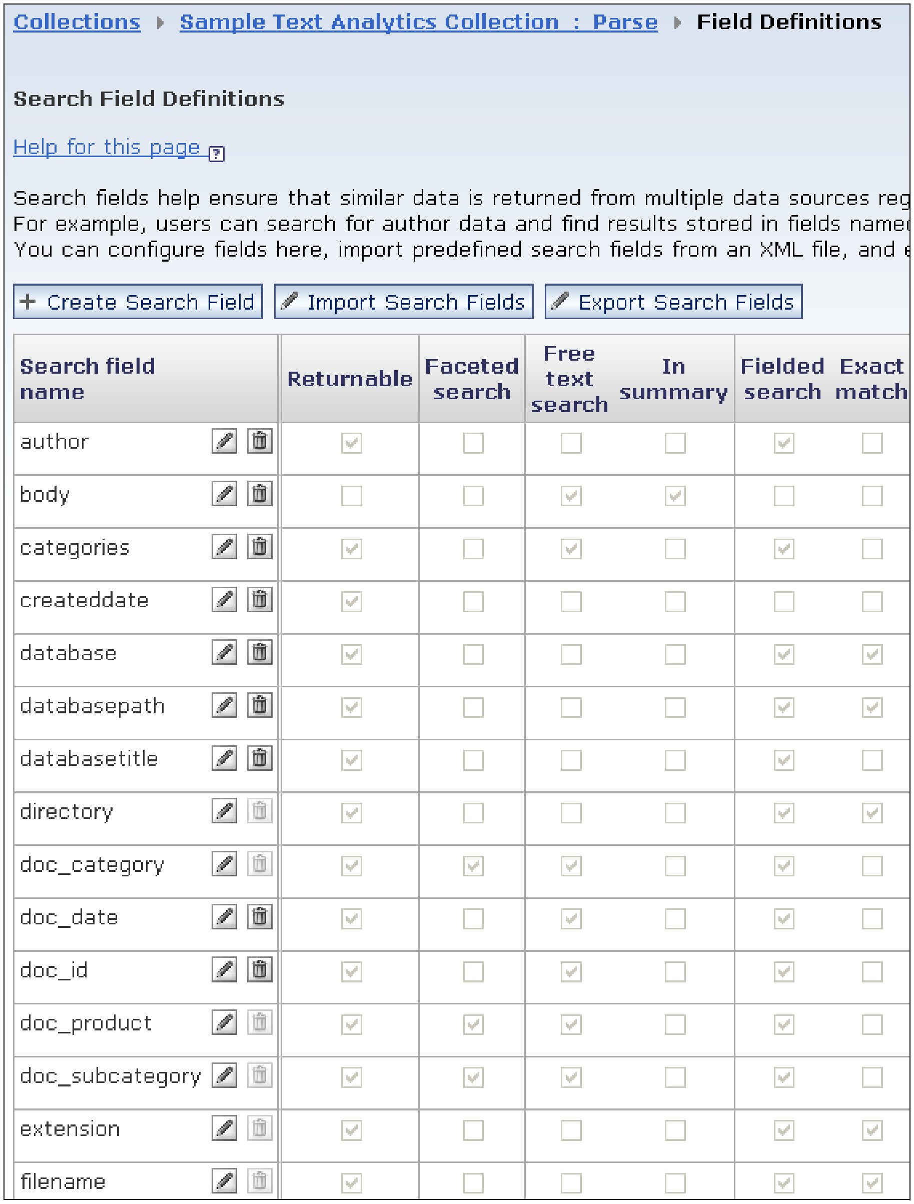
☐ Text sortable
Enables this field to be used for sorting the search results.

☑ Analyzable
Enables this field to be analyzed as document content.

☐ Faceted search
Enables this field to be shown as a facet in the search results.

Figure 4-26 Create a Search Field window

After you create the search fields, the Search Field Definitions window (Figure 4-27) shows the search field definition list.



Collections ▸ Sample Text Analytics Collection : Parse ▸ Field Definitions

Search Field Definitions

Help for this page

Search fields help ensure that similar data is returned from multiple data sources reg
For example, users can search for author data and find results stored in fields name
You can configure fields here, import predefined search fields from an XML file, and e

Search field name	Returnable	Faceted search	Free text search	In summary	Fielded search	Exact match
author	☑	□	□	□	☑	□
body	□	□	☑	☑	□	□
categories	☑	□	☑	□	☑	□
createddate	☑	□	□	□	□	□
database	☑	□	□	□	☑	☑
databasepath	☑	□	□	□	☑	☑
databasetitle	☑	□	□	□	☑	□
directory	☑	□	□	□	☑	☑
doc_category	☑	☑	☑	□	☑	□
doc_date	☑	□	☑	□	☑	□
doc_id	☑	□	☑	□	☑	□
doc_product	☑	☑	☑	□	☑	□
doc_subcategory	☑	☑	☑	□	☑	□
extension	☑	□	□	□	☑	☑
filename	☑	□	□	□	☑	☑

Figure 4-27 Final Search Field Definitions window

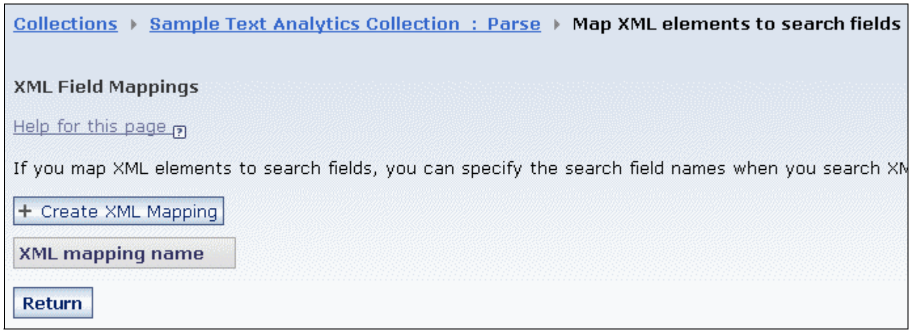
After defining or changing search fields: For field definition changes to take affect, restart the parse and index component if it is already running and redeploy the resources. Additionally, if the documents are already indexed by the component, rebuild the index after restarting the parse and index component.

Mapping the native field to the search field

After you create the search fields, you are ready to map native fields to the search fields based on your design. (See “Mapping native fields to search fields” on page 89.)

To map the native fields to the search fields, follow these steps:

1. From the administration console, go to the **Collections** view. Locate the collection that you want to edit, and click **Edit**.
2. On the **Parse and Index** tab, click **Map XML elements to search fields**.
3. In the XML Field Mappings window (Figure 4-28), click **Create XML Mapping**.


Collections ▸ Sample Text Analytics Collection : Parse ▸ Map XML elements to search fields
XML Field Mappings
Help for this page
+ Create XML Mapping
XML mapping name
Return

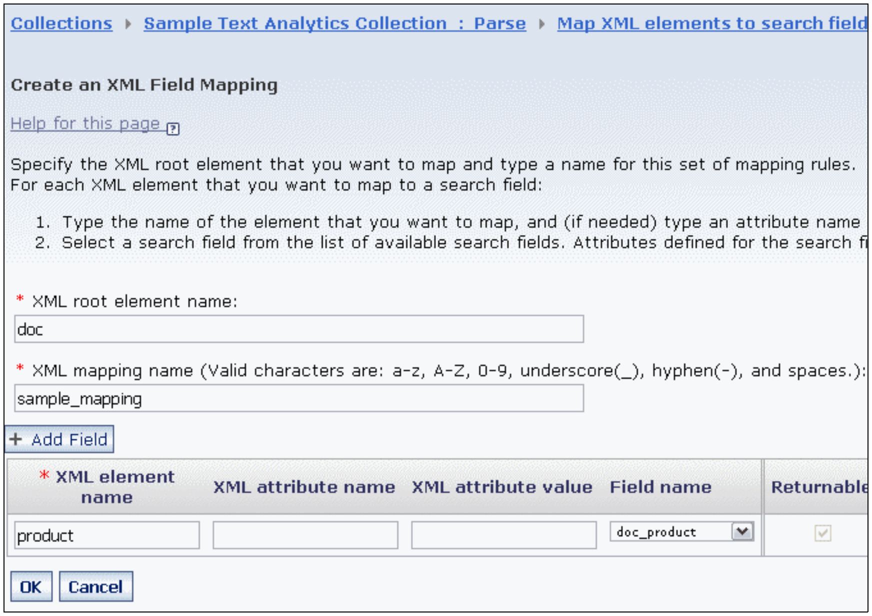
If you map XML elements to search fields, you can specify the search field names when you search XM

Figure 4-28 XML Field Mappings window

4. In the Create an XML Mapping Field window (Figure 4-29), complete the following steps:

   a. Complete the following fields based on the sample data in the test scenario:

      - For XML root element name, enter `doc`.
      - For XML mapping name, enter `sample_mapping`.
      - For XML element name, enter `product`.
      - For Field name, select **doc_product**.



Collections ▸ Sample Text Analytics Collection : Parse ▸ Map XML elements to search field

Create an XML Field Mapping

Help for this page

Specify the XML root element that you want to map and type a name for this set of mapping rules. For each XML element that you want to map to a search field:

1. Type the name of the element that you want to map, and (if needed) type an attribute name
2. Select a search field from the list of available search fields. Attributes defined for the search fi

* XML root element name: doc

* XML mapping name (Valid characters are: a-z, A-Z, 0-9, underscore(_), hyphen(-), and spaces.): sample_mapping

* XML element name	XML attribute name	XML attribute value	Field name	Returnable
product			doc_product	✓

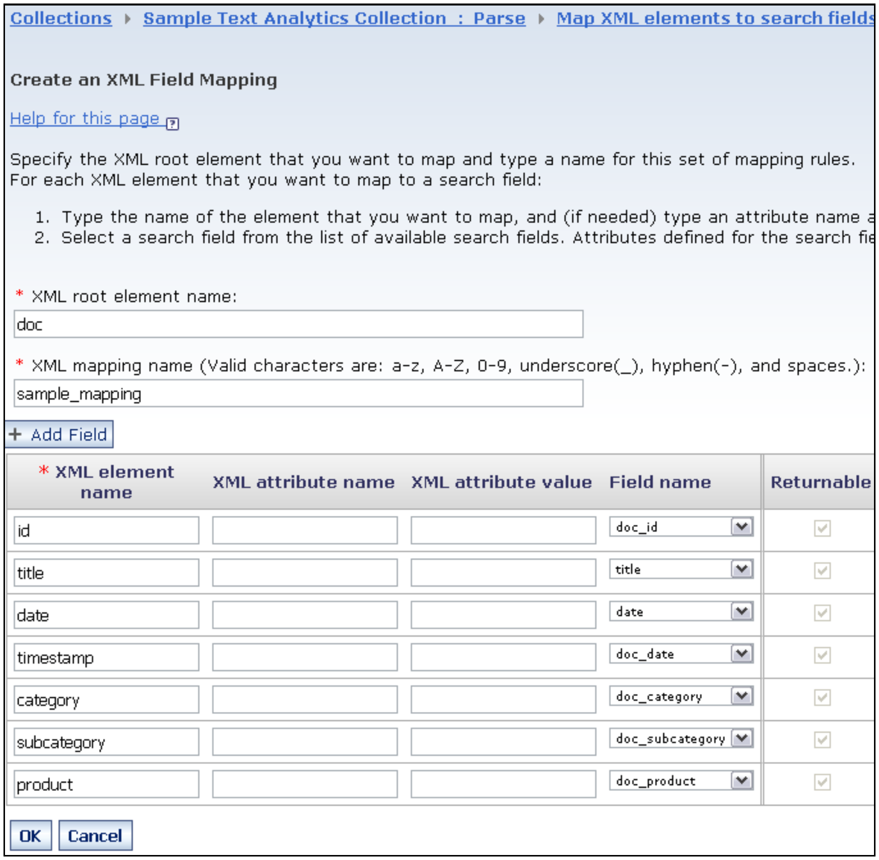
Figure 4-29 Create an XML Field Mapping window

   b. Click **Add Field**, and add the field mappings as shown in Table 4-3.

Table 4-3 Mapping XML element names to field names

XML element name (native field)	Field name (search field)
id	doc_id
title	title
date	date
timestamp	doc_date
category	doc_category
subcategory	doc_subcategory

See Figure 4-30 as a reference.



Collections ▸ Sample Text Analytics Collection : Parse ▸ Map XML elements to search fields

Create an XML Field Mapping

Help for this page

Specify the XML root element that you want to map and type a name for this set of mapping rules. For each XML element that you want to map to a search field:

1. Type the name of the element that you want to map, and (if needed) type an attribute name a
2. Select a search field from the list of available search fields. Attributes defined for the search fie

* XML root element name: doc

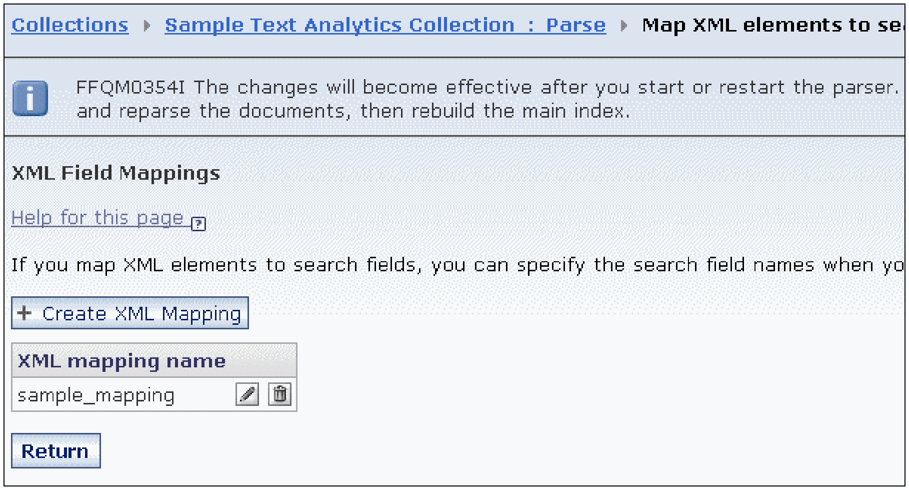
* XML mapping name (Valid characters are: a-z, A-Z, 0-9, underscore(_), hyphen(-), and spaces.): sample_mapping

* XML element name	XML attribute name	XML attribute value	Field name	Returnable
id			doc_id	☑
title			title	☑
date			date	☑
timestamp			doc_date	☑
category			doc_category	☑
subcategory			doc_subcategory	☑
product			doc_product	☑

Figure 4-30 Mapping XML native fields to search fields

c. Click **OK**.

The new XML field mappings are now created as shown in Figure 4-31.


Collections ▸ Sample Text Analytics Collection : Parse ▸ Map XML elements to se

FFQM0354I The changes will become effective after you start or restart the parser. and reparse the documents, then rebuild the main index.

XML Field Mappings

Help for this page

If you map XML elements to search fields, you can specify the search field names when yo

XML mapping name
sample_mapping

Figure 4-31 Sample mapping created

After mapping native fields (XML element names) to search fields (field names): For the field mapping changes to take affect, restart the parse and index component if it is already running and redeploy the resources. Additionally, if the documents are already indexed by the component, rebuild the index after restarting the parse and index component.

Creating facets and mapping search fields to facets

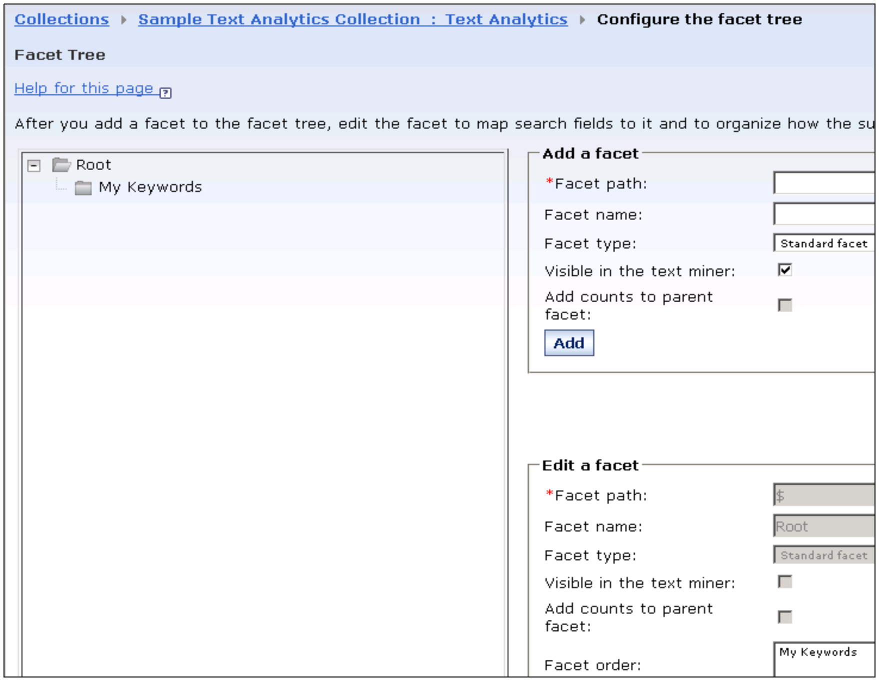
Defining facets in the facet tree consists of two tasks: creating a facet and mapping it to a search field. The facets that you need to create are identified in "Identifying facets and mapping search fields to facets" on page 90.

To create the facets and map search fields to them, follow these steps:

1. From the administration console, go to the **Collections** view. In the list of collections, locate the collection that you want to edit, and click **Edit**.
2. On the **Parse and Index** tab, click **Configure facets**.

3. On the Facet Tree page (Figure 4-32), complete the following steps:

   a. Select the **Root** node from the facet tree.



Collections ▸ Sample Text Analytics Collection : Text Analytics ▸ Configure the facet tree

Facet Tree

Help for this page

After you add a facet to the facet tree, edit the facet to map search fields to it and to organize how the su

- Root
  - My Keywords

Add a facet

*Facet path:

Facet name:

Facet type: Standard facet

Visible in the text miner: ☑

Add counts to parent facet: □

Add

Edit a facet

*Facet path: $

Facet name: Root

Facet type: Standard facet

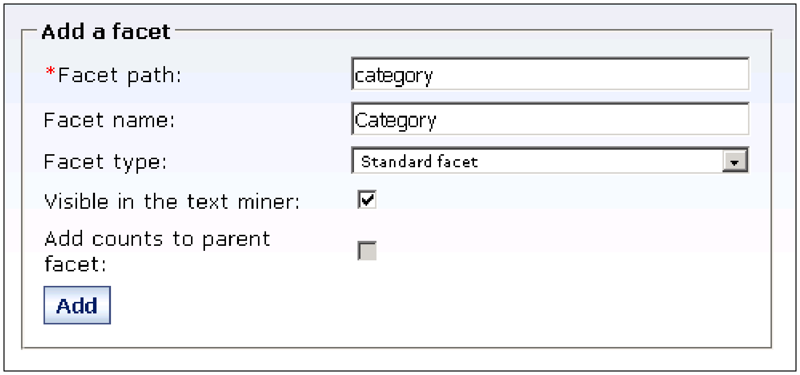
Visible in the text miner: □

Add counts to parent facet: □

Facet order: My Keywords

Figure 4-32 Default facet tree of a collection

   b. In the Add a facet section (Figure 4-33), for the Facet path and Facet name field, type `category`. Select **Standard facet** as the Facet type and click **Add**.



Add a facet

*Facet path: category

Facet name: Category

Facet type: Standard facet

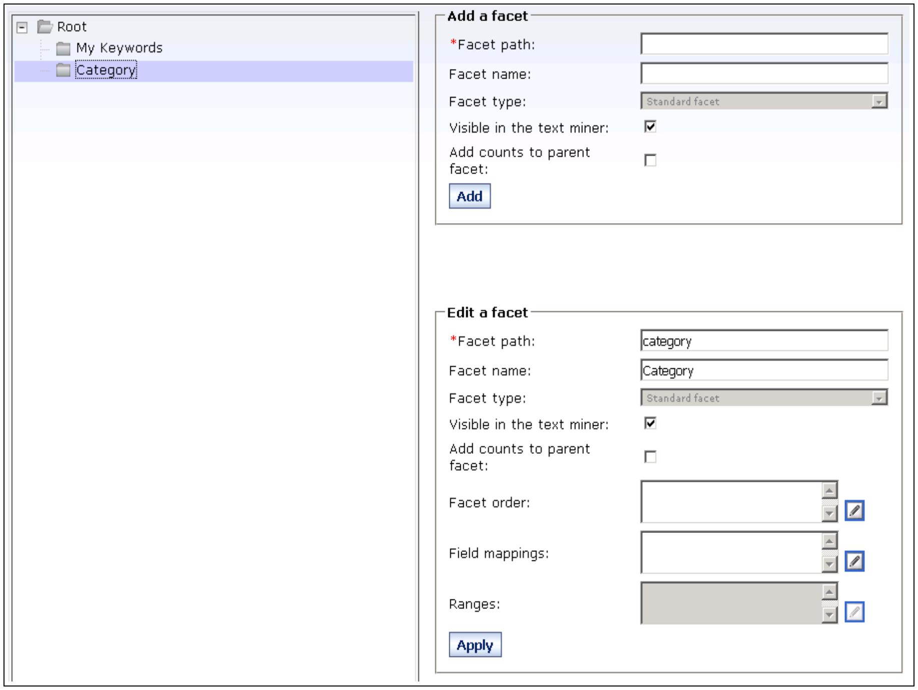
Visible in the text miner: ☑

Add counts to parent facet: □

Add

Figure 4-33 Adding a facet

A new facet is created under the Root node in the Facet tree (Figure 4-34).



Root
My Keywords
Category

Add a facet

*Facet path:
Facet name:
Facet type: Standard facet
Visible in the text miner: ☑
Add counts to parent facet: □

Add

Edit a facet

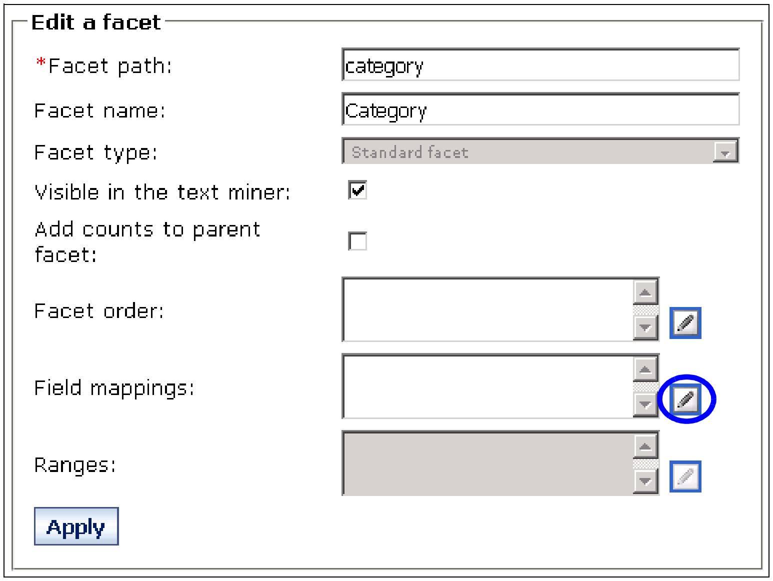
*Facet path: category
Facet name: Category
Facet type: Standard facet
Visible in the text miner: ☑
Add counts to parent facet: □
Facet order:
Field mappings:
Ranges:

Apply

Figure 4-34 New facet added

4. Map the facet to the search field:

   a. Select the **Category** facet that you created in step 3 on page 107.
   b. In the Edit a facet section, for the Field Mappings field, click the **Edit** icon (Figure 4-35).

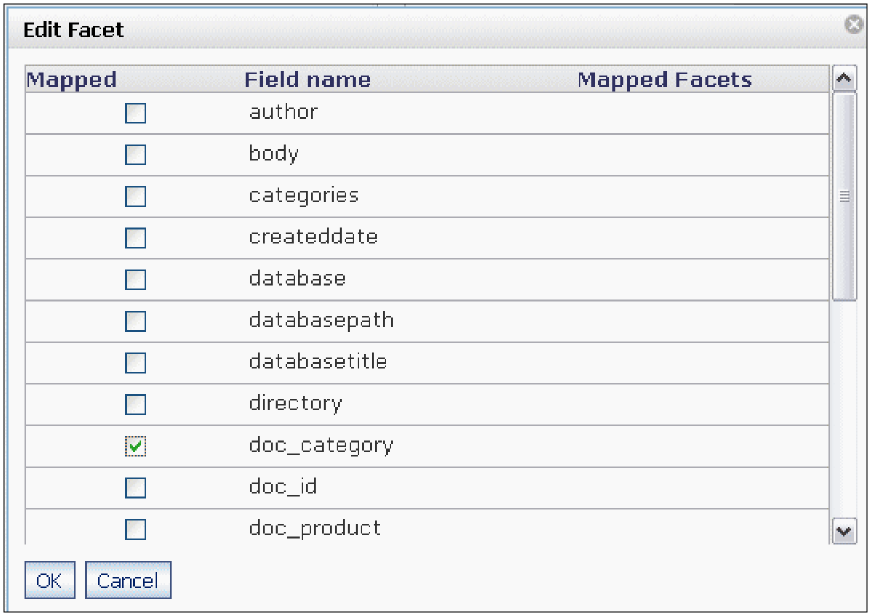


Edit a facet

*Facet path:	category
Facet name:	Category
Facet type:	Standard facet
Visible in the text miner:	☑
Add counts to parent facet:	□
Facet order:	
Field mappings:	
Ranges:	

Figure 4-35 Edit Facet mapping

c. In the Edit Facet window (Figure 4-36), from the list of fields, select **doc_category** and click **OK**.

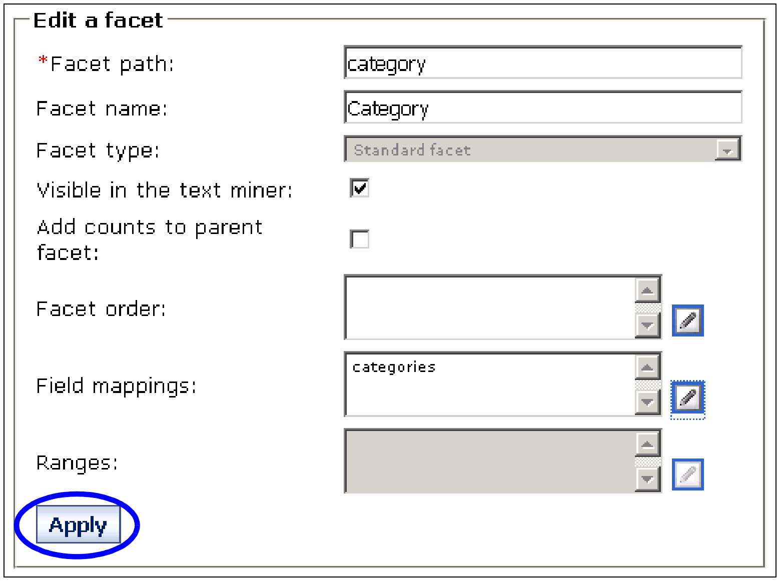


Edit Facet

Mapped	Field name	Mapped Facets
□	author	
□	body	
□	categories	
□	createddate	
□	database	
□	databasepath	
□	databasetitle	
□	directory	
☑	doc_category	
□	doc_id	
□	doc_product	

Figure 4-36 Selecting a search field for mapping to a facet

d. In the Edit a facet section in the Facet Tree window (Figure 4-37), click **Apply**.



Edit a facet

*Facet path: category

Facet name: Category

Facet type: Standard facet

Visible in the text miner: ☑

Add counts to parent facet: □

Facet order:

Field mappings: categories

Ranges:

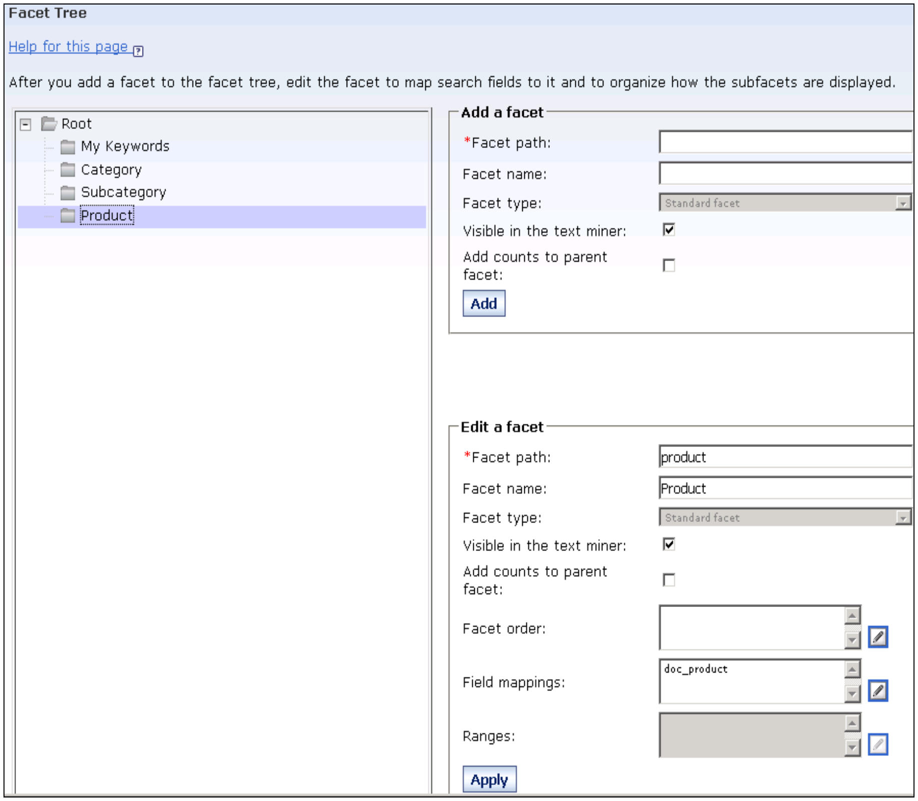
Figure 4-37 Mapping a facet to a search field

5. Repeat step 3 on page 107 through step 4 on page 109, and add facets based on the configuration shown in Table 4-4.

Table 4-4 Facet tree configuration

Facet path	Facet name	Field mappings
subcategory	Subcategory	doc_subcategory
product	Product	doc_product

6. After you create the facets, review your changes in the facet tree (Figure 4-38) and click **OK**.



Facet Tree

Help for this page

After you add a facet to the facet tree, edit the facet to map search fields to it and to organize how the subfacets are displayed.

- Root
  - My Keywords
  - Category
  - Subcategory
  - Product

Add a facet

*Facet path:

Facet name:

Facet type: Standard facet

Visible in the text miner: ☑

Add counts to parent facet: □

Add

Edit a facet

*Facet path: product

Facet name: Product

Facet type: Standard facet

Visible in the text miner: ☑

Add counts to parent facet: □

Facet order:

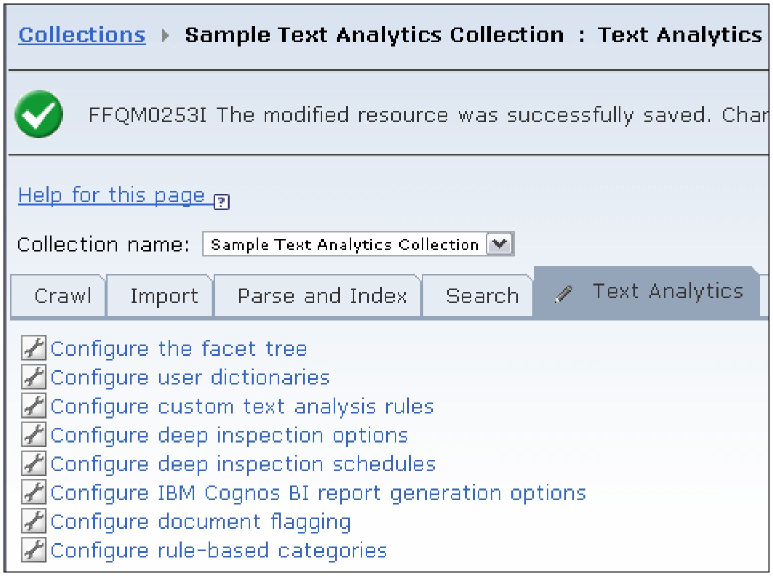
Field mappings: doc_product

Ranges:

Apply

Figure 4-38 Facet tree created

A Confirmation window (Figure 4-39) is displayed after you save the facet tree changes.



Collections ▸ Sample Text Analytics Collection : Text Analytics

FFQM0253I The modified resource was successfully saved. Char

Help for this page

Collection name: Sample Text Analytics Collection

Crawl
Import
Parse and Index
Search
Text Analytics
Configure the facet tree
Configure user dictionaries
Configure custom text analysis rules
Configure deep inspection options
Configure deep inspection schedules
Configure IBM Cognos BI report generation options
Configure document flagging
Configure rule-based categories

Figure 4-39 Confirmation saved for facet tree changes

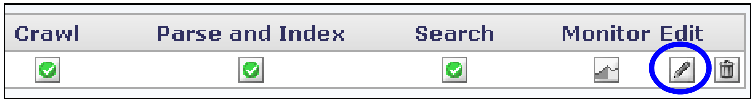
After creating or changing the facets: For the facet tree changes to take affect on already processed data, restart the parse and index component if it is already running and redeploy the resources. Additionally, if the documents are already indexed by the component, rebuild the index after restarting the parse and index component.

Optional: Configuring the date facet

The text miner application displays various graphs to show data over a time period. By default, the date search field is used as the date value for the x-axis in the Time Series, Deviations, and Trend view graphs. The collection can be configured so that you can select a different search field to represent the time period in the views within the text miner application. To use the search field as the date value, you must enable the search field to be searched with a parametric query and the Parametric search option field must be set to Date.

To configure the search fields to be displayed as date options in the text miner application views, follow these steps:

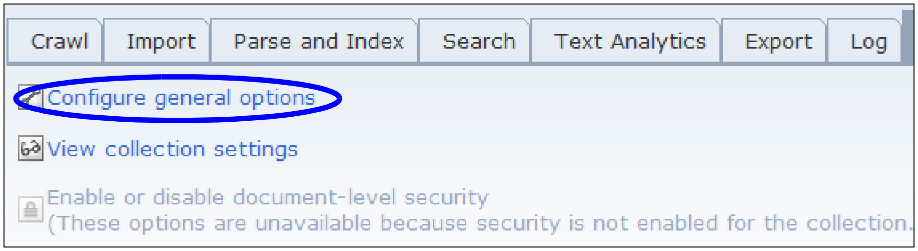
1. From the administration console, on the **Collections** tab (Figure 4-40), in the list of collections, locate the collection that you want to edit, and click **Edit**.



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓		

Figure 4-40 The edit collection icon

2. On the **General** tab (Figure 4-41), click **Configure general options**.



Configure general options

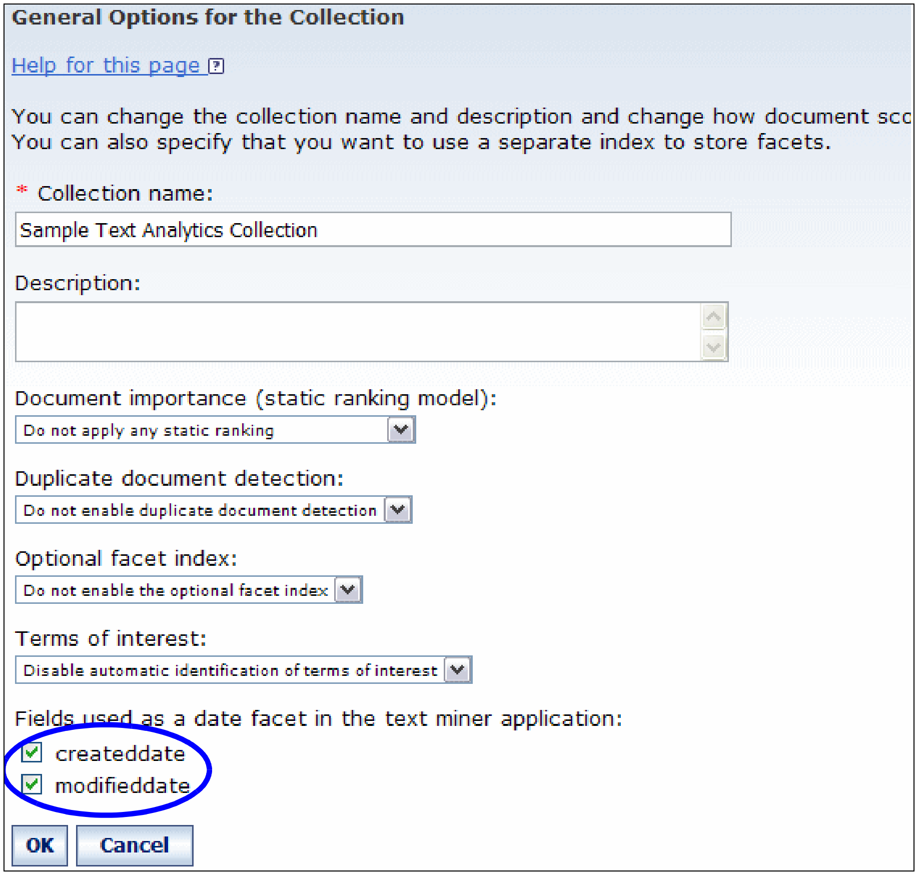
View collection settings

Enable or disable document-level security
(These options are unavailable because security is not enabled for the collection.

Figure 4-41 General tab for the collection

3. For the “Fields used as a date facet in the text miner application” field, select all the fields you want users to use in the text miner views for graphs related to time. Figure 4-42 shows the createdate and modifieddate fields as selected. Click **OK**.

   > **New data options:** The text miner application does not show the new date options until you redeploy the resources.



General Options for the Collection

Help for this page

You can change the collection name and description and change how document sco
You can also specify that you want to use a separate index to store facets.

* Collection name:
Sample Text Analytics Collection

Description:

Document importance (static ranking model):
Do not apply any static ranking

Duplicate document detection:
Do not enable duplicate document detection

Optional facet index:
Do not enable the optional facet index

Terms of interest:
Disable automatic identification of terms of interest

Fields used as a date facet in the text miner application:

☑ createddate

☑ modifieddate

Figure 4-42 Configure general options for the collection

To learn more about handling of the date facet within the text miner application, see “Changing the Date facet” on page 230.

Optional: Creating a range facet and defining ranges

In the previous section, you created standard facets. You can also create a decimal range facet or a date range facet. The range facets are useful to analyze the data within a certain range. This section briefly explains how you can set up the range facet.

In this section, you see how to set up the date range facet. This example is not used in the scenario in Chapter 6, “Text miner application: Views” on page 217.

Defining range facets in the facet tree consists of three tasks:

1. Creating a facet as a range facet
2. Mapping the facet to a decimal
3. Setting the search field as date parametric search and defining a range for the range facet

To define the range facets, follow these steps:

1. From the administration console, click the **Collections** tab. In the list of collections, locate the collection that you want to edit, and click **Edit** (Figure 4-43).

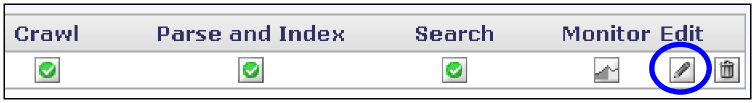


Figure 4-43 Edit a collection

2. On the **Parse and Index** tab, click **Configure facets**.
3. On the Facet Tree page, complete these steps:
   a. Select the **Root** node from the facet tree.

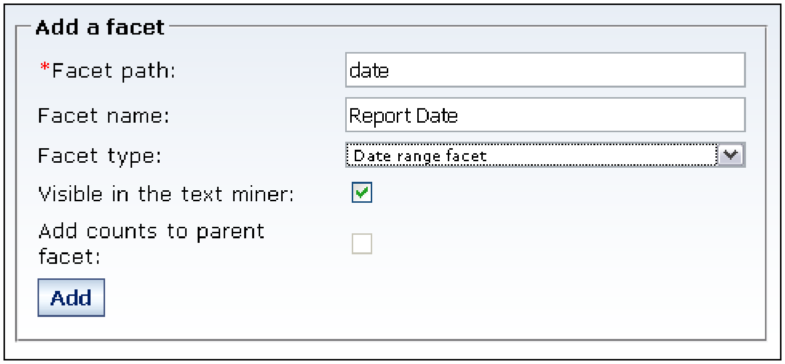
> **Creating a range facet:** The decimal range facet or the date range facet can be created just under the **Root** node.

b. In the Add a facet section (Figure 4-44), complete the following steps:

i. For the Facet path, type `date`.
ii. For the Facet name field, type `Report Date`.
iii. For the Facet type field, select **Date range facet**.

> **Facet type:** When you create a decimal range facet, select **Decimal range facet** for the Facet type field.

iv. Click **Add**.



Add a facet

*Facet path: date

Facet name: Report Date

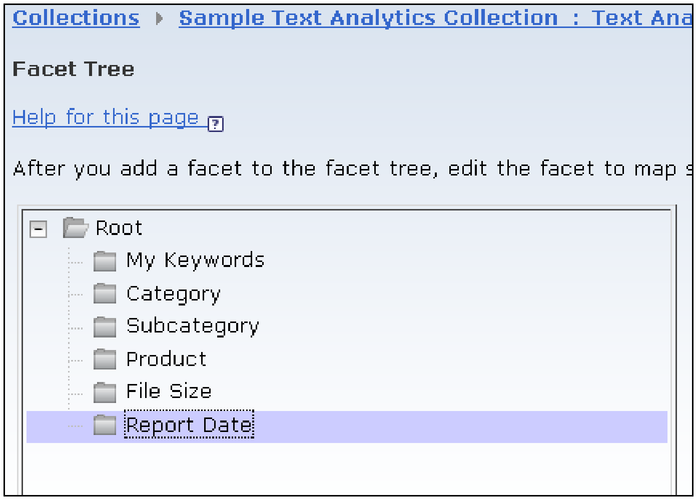
Facet type: Date range facet

Visible in the text miner: ☑

Add counts to parent facet: □

Figure 4-44 Adding date range facet

A new facet is created under the Root node in the Facet tree called Report Date (Figure 4-45).



Facet Tree

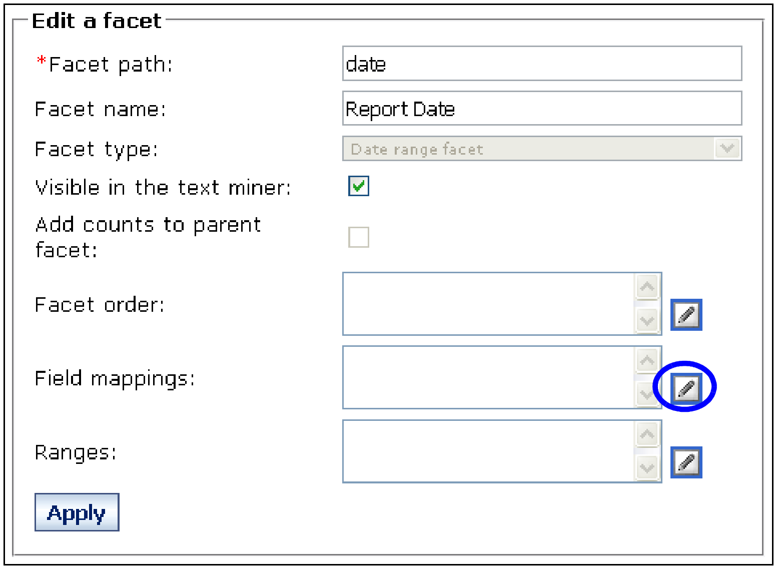
After you add a facet to the facet tree, edit the facet to map s

Figure 4-45 Facet tree

4. Map the facet to the parametric search field:

   a. Select the date range facet or decimal range facet that you created in step 3 on page 116.

   b. In the Edit a facet section (Figure 4-46), for the Field mappings field, click the **Edit** icon.

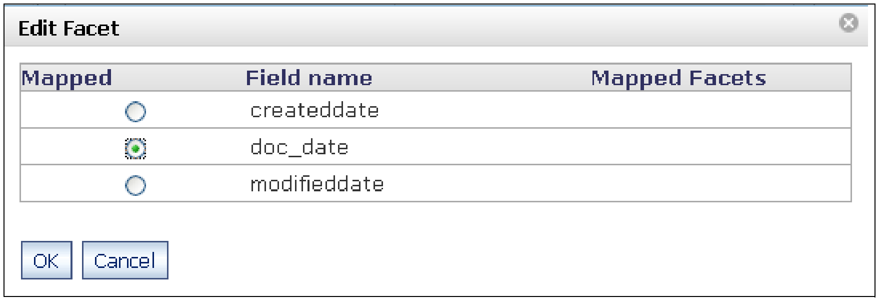


Edit a facet

*Facet path:	date
Facet name:	Report Date
Facet type:	Date range facet
Visible in the text miner:	☑
Add counts to parent facet:	□
Facet order:	
Field mappings:	
Ranges:	

Figure 4-46 Edit field mappings

   c. In the Edit Facet window, select the date parametric field for the date range facet or the decimal parametric search field, and click **OK** (Figure 4-47).



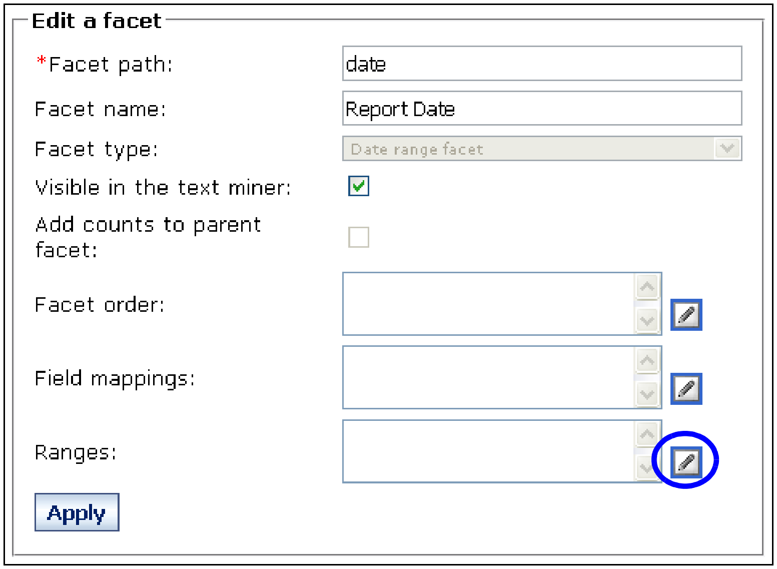
Edit Facet

Mapped	Field name	Mapped Facets
○	createddate	
●	doc_date	
○	modifieddate	

Figure 4-47 List of parametric date type fields

Selecting a field: When selecting the field for range facet mapping, only the parametric search fields are displayed. If you do not see the search field that you want to configure, the search field might be configured as date parametric search field or decimal parametric search field.

5. Define the ranges for the range facet:
   a. Select the range facet that you created in step 3 on page 116.
   b. In the Edit a facet section (Figure 4-48), click the **Edit** icon next to the Ranges field.



Edit a facet

*Facet path: date

Facet name: Report Date

Facet type: Date range facet

Visible in the text miner: ☑

Add counts to parent facet: □

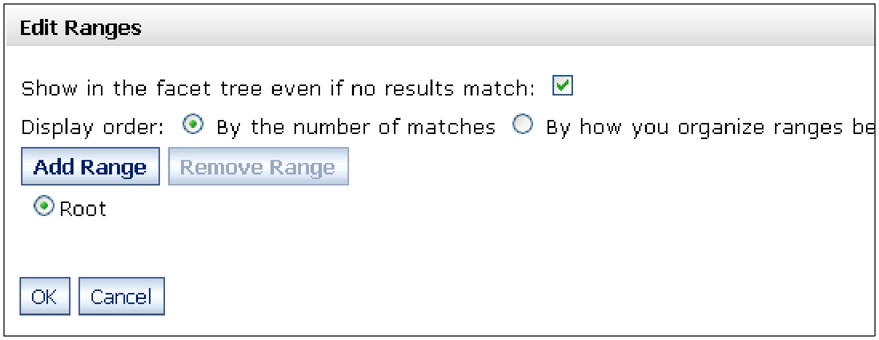
Facet order:

Field mappings:

Ranges:

Figure 4-48 Edit ranges

c. In the Edit Ranges window (Figure 4-49), select the **Root** radio button and click **Add Range** to add the range.



Edit Ranges

Show in the facet tree even if no results match: ☑

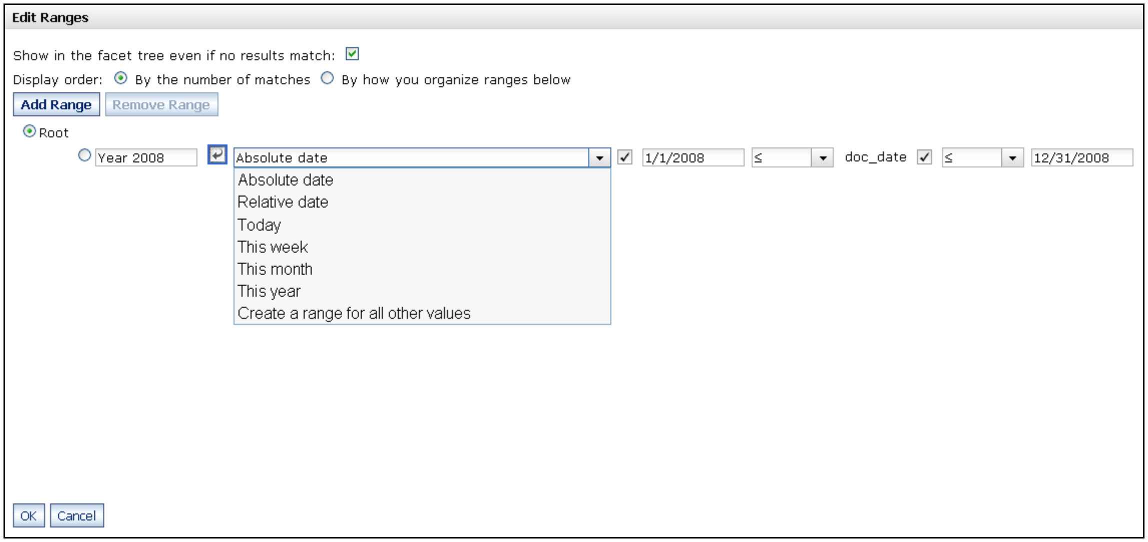
Display order: ● By the number of matches ○ By how you organize ranges be

● Root

Figure 4-49 Creating the range

Now the new range list is displayed.

d. Select the range type and range name that you want to use. For the date range facet, several range types are available (Figure 4-50).



Edit Ranges

Show in the facet tree even if no results match: ☑

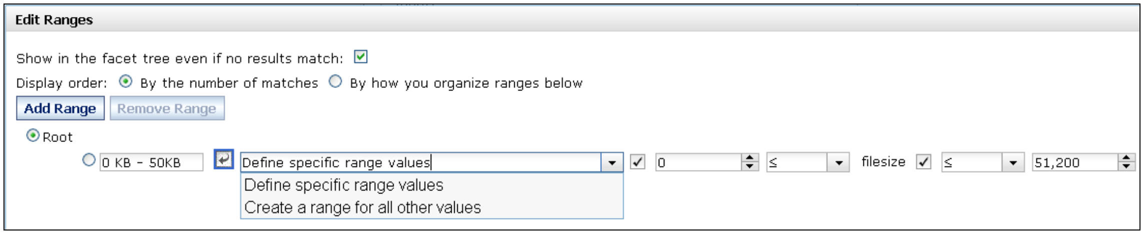
Display order: ● By the number of matches ○ By how you organize ranges below

● Root

○ Year 2008 ☑ Absolute date ☑ 1/1/2008 ≤ doc_date ☑ ≤ 12/31/2008

Figure 4-50 Defining the range for the date range facet

For the decimal range facet, two range types are available (Figure 4-51).



Edit Ranges

Show in the facet tree even if no results match: ☑

Display order: ● By the number of matches ○ By how you organize ranges below

Add Range | Remove Range

● Root

○ 0 KB - 50KB	Define specific range values	☑	0	≤	filesize	☑ ≤	51,200

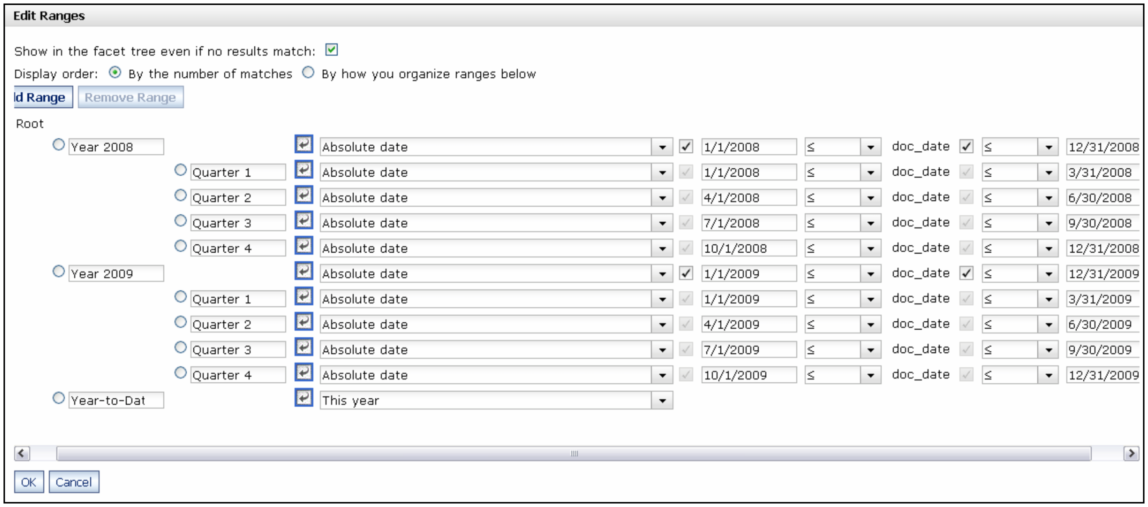
Define specific range values
Create a range for all other values

Figure 4-51 Defining the range for the decimal range facet

After you finish adding all ranges that you want to add, click **OK**.

> **Defining ranges within a range:** As shown in Figure 4-52, you can define multiple ranges within a range.

6. Repeat step 3 on page 116 through step 5 on page 119. Add ranges based on your requirements. Figure 4-52 shows a set of ranges and subranges defined for illustration purpose.



Edit Ranges

Show in the facet tree even if no results match: ☑

Display order: ● By the number of matches ○ By how you organize ranges below

ld Range | Remove Range

Root

○ Year 2008		Absolute date	☑	1/1/2008	≤	doc_date	☑ ≤	12/31/2008
	○ Quarter 1	Absolute date	☑	1/1/2008	≤	doc_date	☑ ≤	3/31/2008
	○ Quarter 2	Absolute date	☑	4/1/2008	≤	doc_date	☑ ≤	6/30/2008
	○ Quarter 3	Absolute date	☑	7/1/2008	≤	doc_date	☑ ≤	9/30/2008
	○ Quarter 4	Absolute date	☑	10/1/2008	≤	doc_date	☑ ≤	12/31/2008
○ Year 2009		Absolute date	☑	1/1/2009	≤	doc_date	☑ ≤	12/31/2009
	○ Quarter 1	Absolute date	☑	1/1/2009	≤	doc_date	☑ ≤	3/31/2009
	○ Quarter 2	Absolute date	☑	4/1/2009	≤	doc_date	☑ ≤	6/30/2009
	○ Quarter 3	Absolute date	☑	7/1/2009	≤	doc_date	☑ ≤	9/30/2009
	○ Quarter 4	Absolute date	☑	10/1/2009	≤	doc_date	☑ ≤	12/31/2009
○ Year-to-Dat		This year						

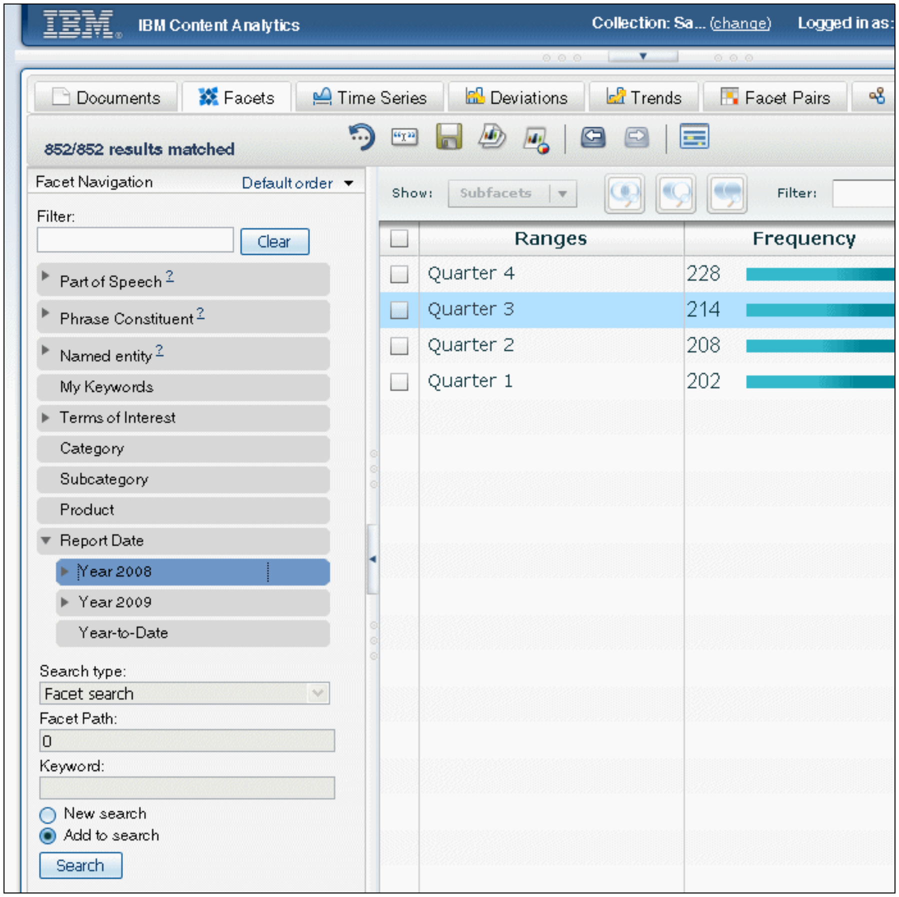
OK | Cancel

Figure 4-52 Ranges for date type facet

7. After you finish defining the ranges, review your changes in the facet tree and click **OK**.

After defining or changing the range facets: After creating the range facet, restart the parse and index component if it is already running. Also redeploy the resources in order for the facet tree changes to take affect on already processed data. Additionally, if the documents are already indexed by the component, rebuild the index after restarting the parse and index component. However, when you add or change a range definition in the existing facet, it is effective immediately in the text miner application and does not require restarting the parse and index component or redeploying the resources.

Figure 4-53 shows an example of how a range facet, Report Date, is displayed in the text miner application.



Search type: Facet search

Facet Path: 0

Keyword:

○ New search
● Add to search

□	Ranges	Frequency
□	Quarter 4	228
□	Quarter 3	214
□	Quarter 2	208
□	Quarter 1	202

Figure 4-53 Range facet in the text miner application

For more information about creating and configuring range facets, go to the IBM Content Analytics Information Center available at the following address, and search on *configuring the facet tree for text analytics collections*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

4.3.3 Defining and configuring a crawler

A crawler discovers the information in the source repository and crawls the data when you start it. Content Analytics provides multiple crawlers for different types of enterprise data sources. Depending on the type of data source, the configuration for creating crawlers varies. Documenting the crawler creation instructions for all the supported data sources is outside the scope of this book.

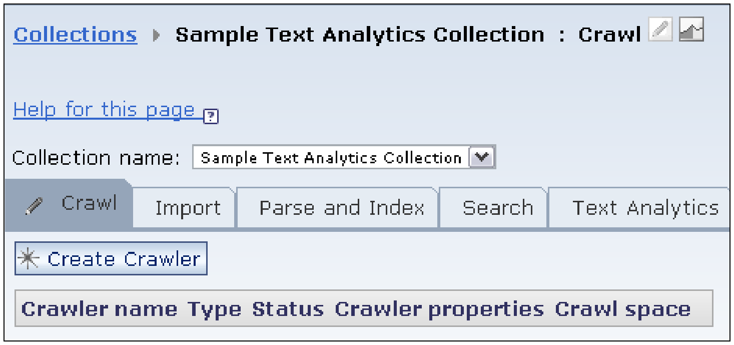
This section concentrates on only one crawler, the *file system crawler*, which crawls the sample data provided by Content Analytics. You learn how to perform the following crawler-related tasks:

- Creating a crawler
- Optional: Scheduling a crawler
- Starting the crawler component
- Stopping the crawler component
- Checking the crawler component status
- Creating a custom crawler plug-in

Creating a crawler

To create a windows file system crawler, follow these steps:

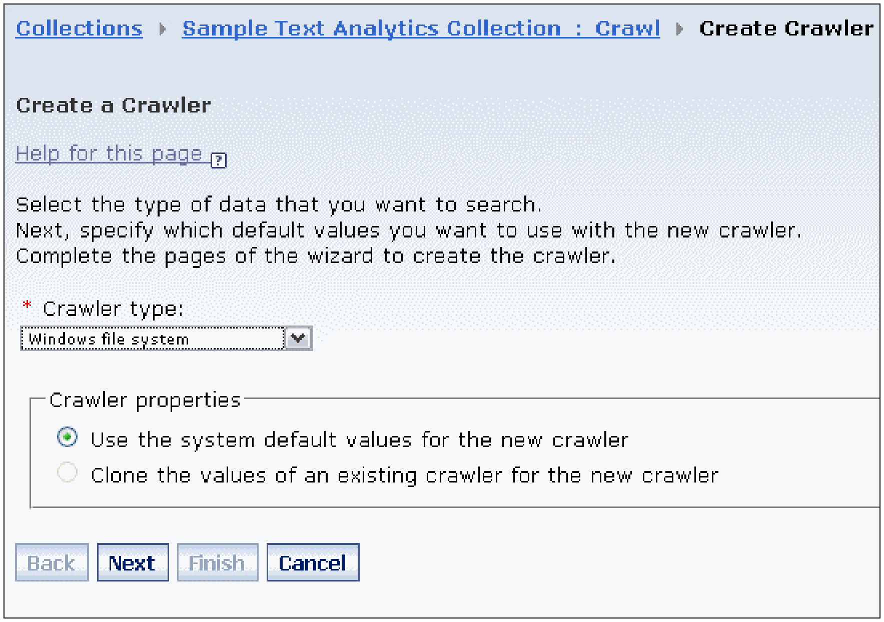
1. From the administration console, go to the **Collections** view. Locate the collection that you want to edit, and click **Edit**.
2. On the **Crawl** tab, click the **Create Crawler** button (Figure 4-54).



Collection name: Sample Text Analytics Collection

Figure 4-54 Crawl tab in edit mode to create a crawler

3. In the Create a crawler window (Figure 4-55), for Crawler type, select **Windows file system** and click **Next**.


Collections ▸ Sample Text Analytics Collection : Crawl ▸ Create Crawler

Create a Crawler

Select the type of data that you want to search.
Next, specify which default values you want to use with the new crawler.
Complete the pages of the wizard to create the crawler.

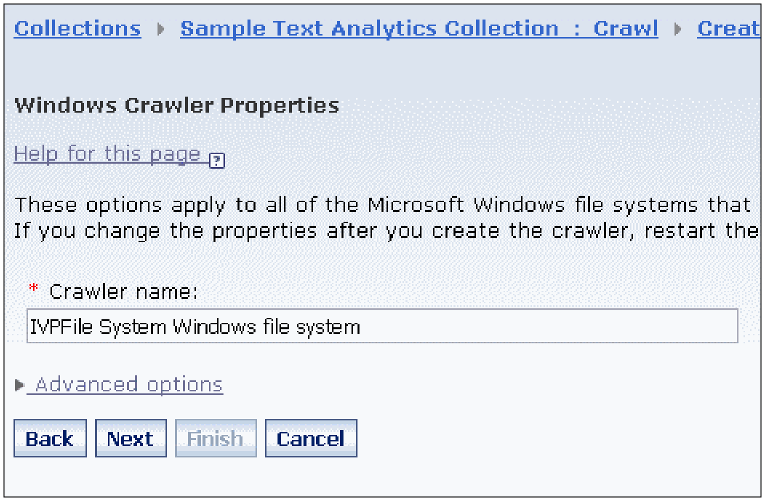
* Crawler type:
Windows file system

Crawler properties

● Use the system default values for the new crawler
○ Clone the values of an existing crawler for the new crawler

Figure 4-55 Choosing a crawler type

4. In the Windows Crawler Properties window (Figure 4-56), for Crawler name, type `IVPFile System Windows file system` and click **Next**.


Collections ▸ Sample Text Analytics Collection : Crawl ▸ Creat

Windows Crawler Properties

These options apply to all of the Microsoft Windows file systems that
If you change the properties after you create the crawler, restart the

* Crawler name:
IVPFile System Windows file system

Advanced options
Back
Next
Finish
Cancel

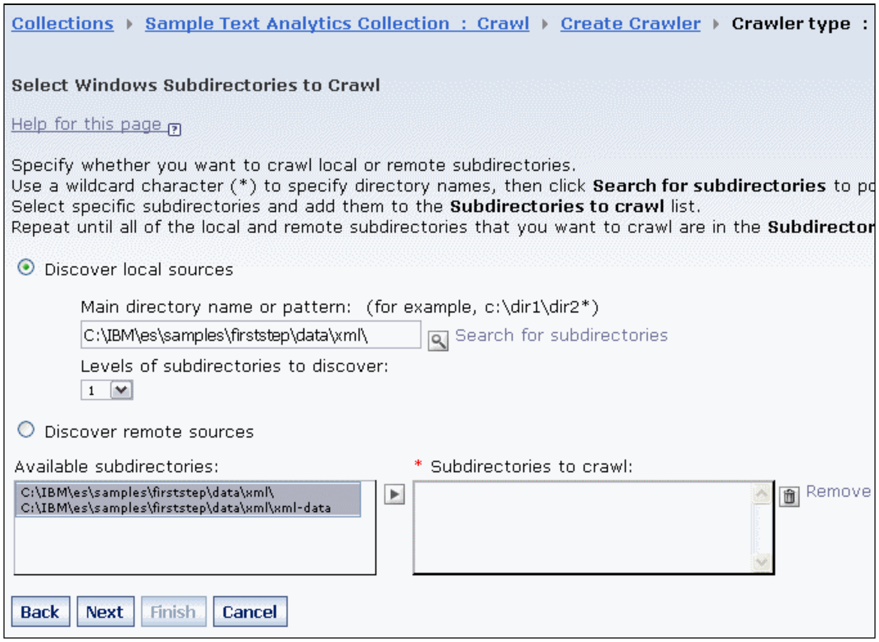
Figure 4-56 Entering the crawler name

5. In the Select Windows Subdirectories to Crawl window, complete these steps:

   a. For the root directory name, enter the path to the sample XML data as `C:\IBM\es\samples\firststep\data\xml\xml-data`.

   b. Click **Search for subdirectories**.

   c. From the Available subdirectories box, select **C:\IBM\es\samples\firststep\data\xml\xml-data** and click the arrow to move it to the Subdirectories to crawl box (Figure 4-57).



Collections ▸ Sample Text Analytics Collection : Crawl ▸ Create Crawler ▸ Crawler type :

Select Windows Subdirectories to Crawl

Help for this page

Specify whether you want to crawl local or remote subdirectories.
Use a wildcard character (*) to specify directory names, then click **Search for subdirectories** to po
Select specific subdirectories and add them to the **Subdirectories to crawl** list.
Repeat until all of the local and remote subdirectories that you want to crawl are in the **Subdirector**

● Discover local sources

Main directory name or pattern: (for example, c:\dir1\dir2*)

C:\IBM\es\samples\firststep\data\xml\

Search for subdirectories

Levels of subdirectories to discover:

1

○ Discover remote sources

Available subdirectories:

C:\IBM\es\samples\firststep\data\xml\
C:\IBM\es\samples\firststep\data\xml\xml-data

* Subdirectories to crawl:

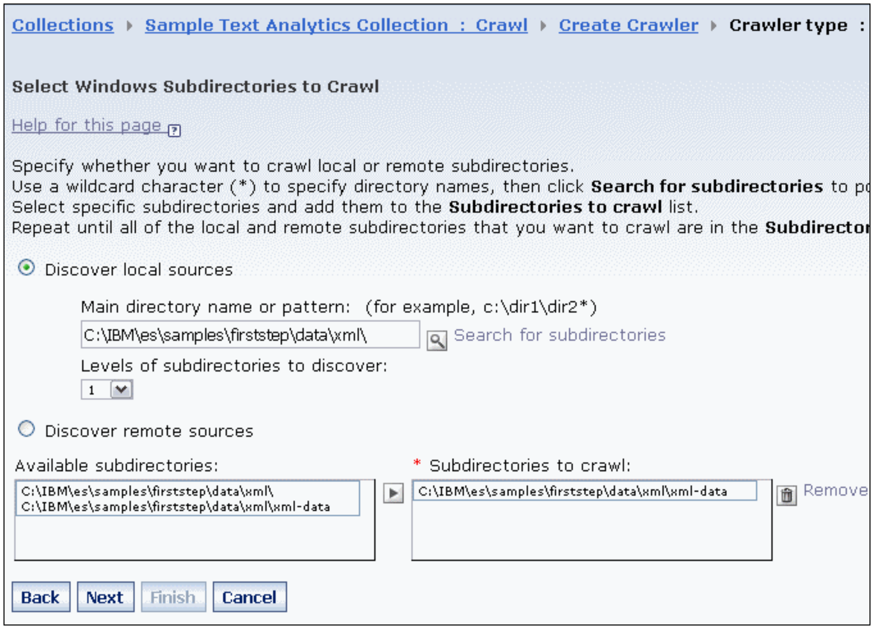
Remove

Back | Next | Finish | Cancel

Figure 4-57 Selecting a subdirectory to crawl

Figure 4-58 shows the completed window.

d. Click **Next**.



Collections ▸ Sample Text Analytics Collection : Crawl ▸ Create Crawler ▸ Crawler type :

Select Windows Subdirectories to Crawl

Help for this page

Specify whether you want to crawl local or remote subdirectories.
Use a wildcard character (*) to specify directory names, then click **Search for subdirectories** to po
Select specific subdirectories and add them to the **Subdirectories to crawl** list.
Repeat until all of the local and remote subdirectories that you want to crawl are in the **Subdirector**

● Discover local sources

Main directory name or pattern: (for example, c:\dir1\dir2*)
C:\IBM\es\samples\firststep\data\xml\ Search for subdirectories
Levels of subdirectories to discover:
1

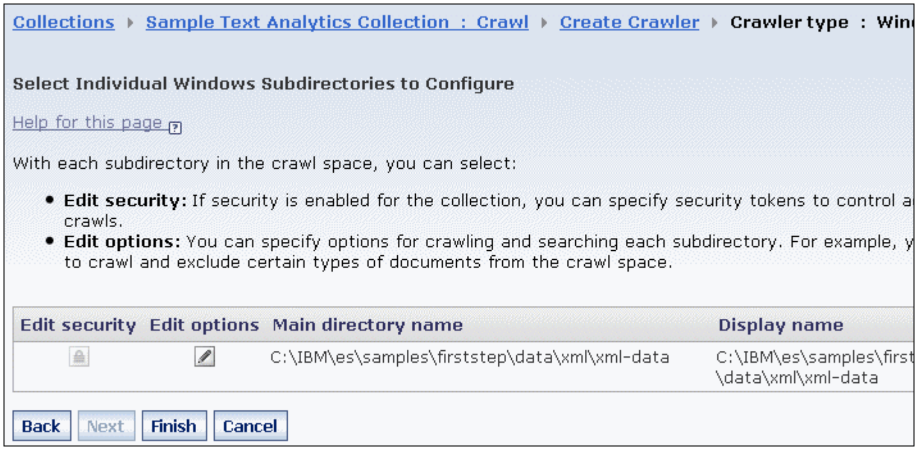
○ Discover remote sources

Available subdirectories:
C:\IBM\es\samples\firststep\data\xml\
C:\IBM\es\samples\firststep\data\xml\xml-data

* Subdirectories to crawl:
C:\IBM\es\samples\firststep\data\xml\xml-data

Figure 4-58 File system crawler configuration for the use-case scenario

6. In the Select Individual Windows Subdirectories to Configure window (Figure 4-59), click **Finish**.



Collections ▸ Sample Text Analytics Collection : Crawl ▸ Create Crawler ▸ Crawler type : Win

Select Individual Windows Subdirectories to Configure

Help for this page

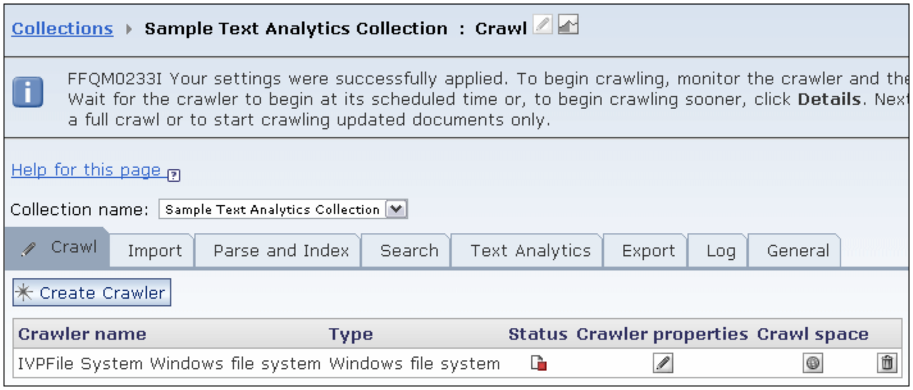
With each subdirectory in the crawl space, you can select:

- **Edit security:** If security is enabled for the collection, you can specify security tokens to control a crawls.
- **Edit options:** You can specify options for crawling and searching each subdirectory. For example, y to crawl and exclude certain types of documents from the crawl space.

Edit security	Edit options	Main directory name	Display name
🔒	✎	C:\IBM\es\samples\firststep\data\xml\xml-data	C:\IBM\es\samples\first \data\xml\xml-data

Figure 4-59 Finishing the crawler creation

Figure 4-60 shows the confirmation message that is displayed after creating a crawler.



Collections ▸ Sample Text Analytics Collection : Crawl

FFQM0233I Your settings were successfully applied. To begin crawling, monitor the crawler and the Wait for the crawler to begin at its scheduled time or, to begin crawling sooner, click **Details**. Next a full crawl or to start crawling updated documents only.

Help for this page

Collection name: Sample Text Analytics Collection

Crawler name	Type	Status	Crawler properties	Crawl space	
IVPFile System Windows file system	Windows file system				

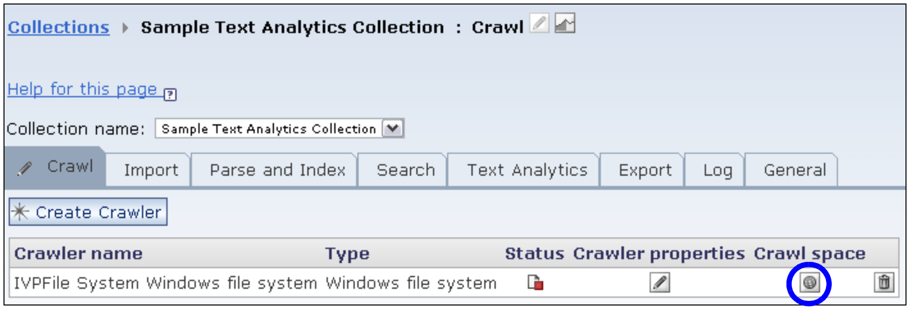
Figure 4-60 Crawler created

Optional: Scheduling a crawler

By default, a crawler starts crawling data as soon as you start the crawler component. After you create a crawler, you can set a schedule for the crawler to run at a specific day or date and time. With Content Analytics, you can set two different schedules for when the crawler starts. In addition, for each schedule, you can select a different crawl type option. For example, you can set a schedule to crawl only new and modified data once a day on weekdays to get daily updates and set another schedule to recrawl all the documents at off-peak times on weekends.

To set a schedule for crawler after it is created, follow these steps:

1. From the administration console, go to the Collections view. Locate the collector you want to work with and click **Edit**.
2. On the **Crawl** tab, click the **Crawl Space** icon (Figure 4-61).



Collections ▸ Sample Text Analytics Collection : Crawl

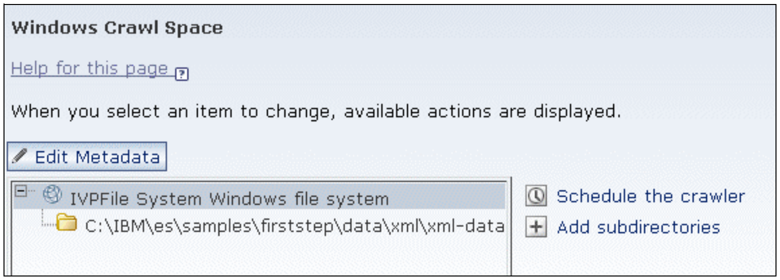
Help for this page

Collection name: Sample Text Analytics Collection

Crawler name	Type	Status	Crawler properties	Crawl space	
IVPFile System Windows file system	Windows file system				

Figure 4-61 Crawl tab in edit mode

3. In the Windows Crawl Space window (Figure 4-62), click **Schedule the crawler**.



Windows Crawl Space

Help for this page

When you select an item to change, available actions are displayed.

Edit Metadata

IVPFile System Windows file system

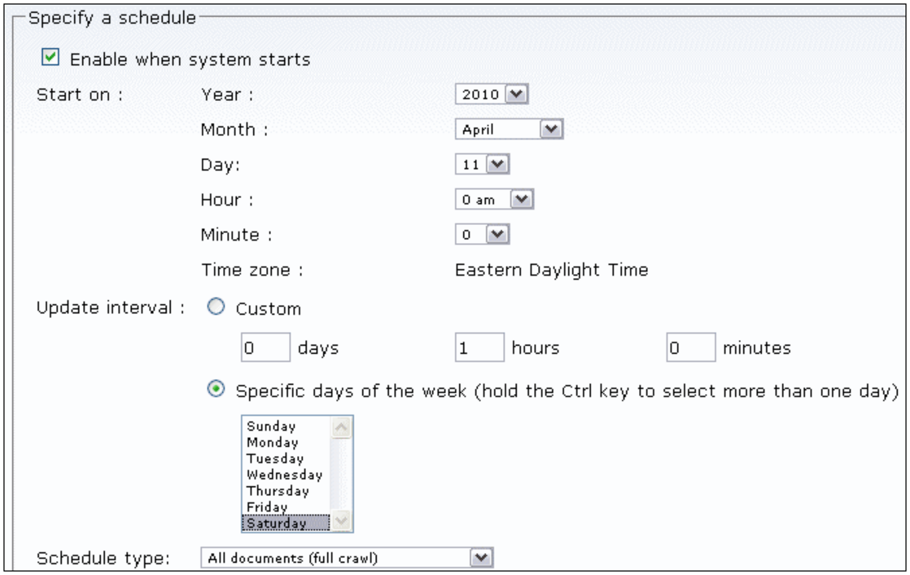
C:\IBM\es\samples\firststep\data\xml\xml-data

Schedule the crawler

Add subdirectories

Figure 4-62 Windows Crawl space window

4. In the Specify a schedule window, set the appropriate setting that is suitable for your use case. Figure 4-63 shows an example of crawling all documents every Saturday at midnight.



Specify a schedule

☑ Enable when system starts

Start on : Year : 2010
Month : April
Day: 11
Hour : 0 am
Minute : 0
Time zone : Eastern Daylight Time

Update interval : ○ Custom
0 days 1 hours 0 minutes

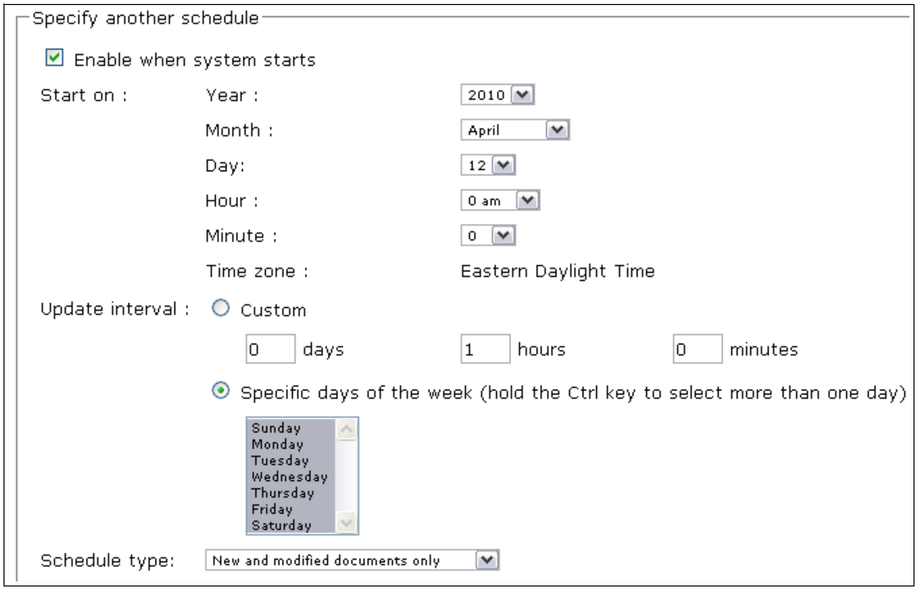
● Specific days of the week (hold the Ctrl key to select more than one day)

Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

Schedule type: All documents (full crawl)

Figure 4-63 Weekly crawl of all documents schedule

Figure 4-64 shows an example of crawling new and modified documents everyday at midnight.



Specify another schedule

☑ Enable when system starts

Start on : Year : 2010

Month : April

Day: 12

Hour : 0 am

Minute : 0

Time zone : Eastern Daylight Time

Update interval : ○ Custom

0 days 1 hours 0 minutes

● Specific days of the week (hold the Ctrl key to select more than one day)

Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

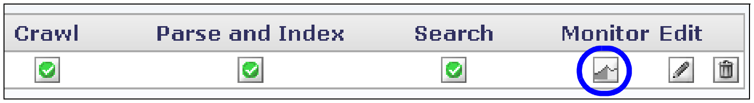
Schedule type: New and modified documents only

Figure 4-64 Daily crawl schedule of new and modified documents

Starting the crawler component

To start the crawler from the administration console, follow these steps:

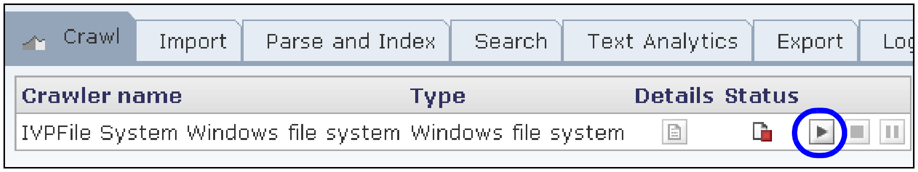
1. From the Collections view, click the **Monitor** icon (Figure 4-65).



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓		

Figure 4-65 Collections view with editing and monitoring options

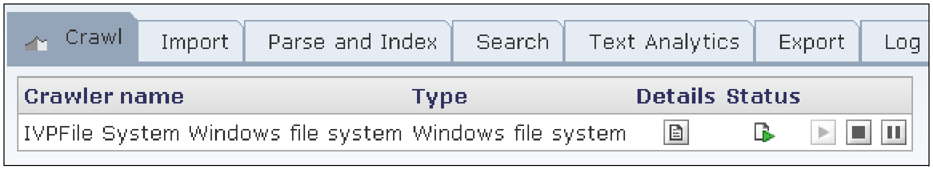
2. Click the **Start** icon (Figure 4-66) to start the crawler.



Crawler name	Type	Details	Status
IVPFile System Windows file system	Windows file system		

Figure 4-66 Crawler in stop node showing the Start icon

Figure 4-67 shows the status when the crawler is started.



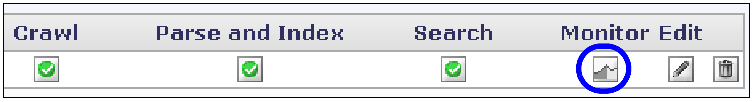
Crawler name	Type	Details	Status
IVPFile System Windows file system	Windows file system		

Figure 4-67 The file system crawler in start status

Stopping the crawler component

To stop the crawler using the administration console, follow these steps:

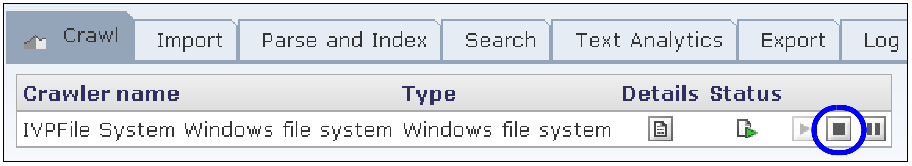
1. From the Collections view, click the **Monitor** icon (Figure 4-68) associated with your collection.



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓		

Figure 4-68 Monitor icon in the collection view

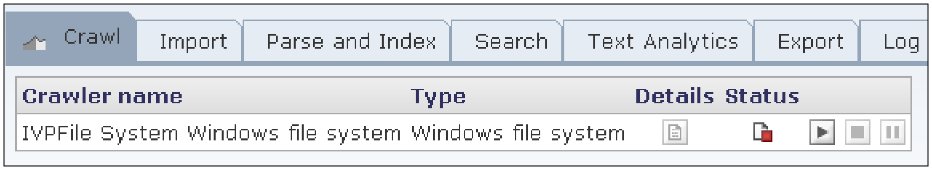
2. Click the **Crawl** tab and click the **Stop** icon (Figure 4-69) to stop the crawler.



Crawler name	Type	Details	Status
IVPFile System Windows file system	Windows file system		

Figure 4-69 Crawler in start mode

Figure 4-70 shows the status when the crawler is stopped.

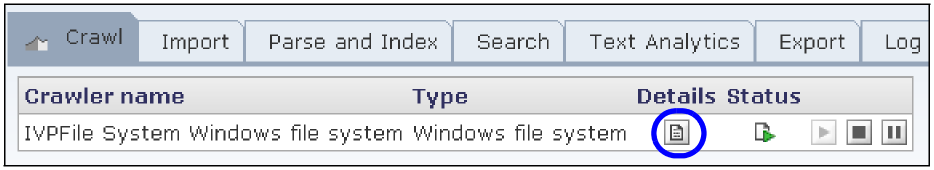


Crawler name	Type	Details	Status
IVPFile System Windows file system	Windows file system		

Figure 4-70 Crawler in stop mode

Checking the crawler component status

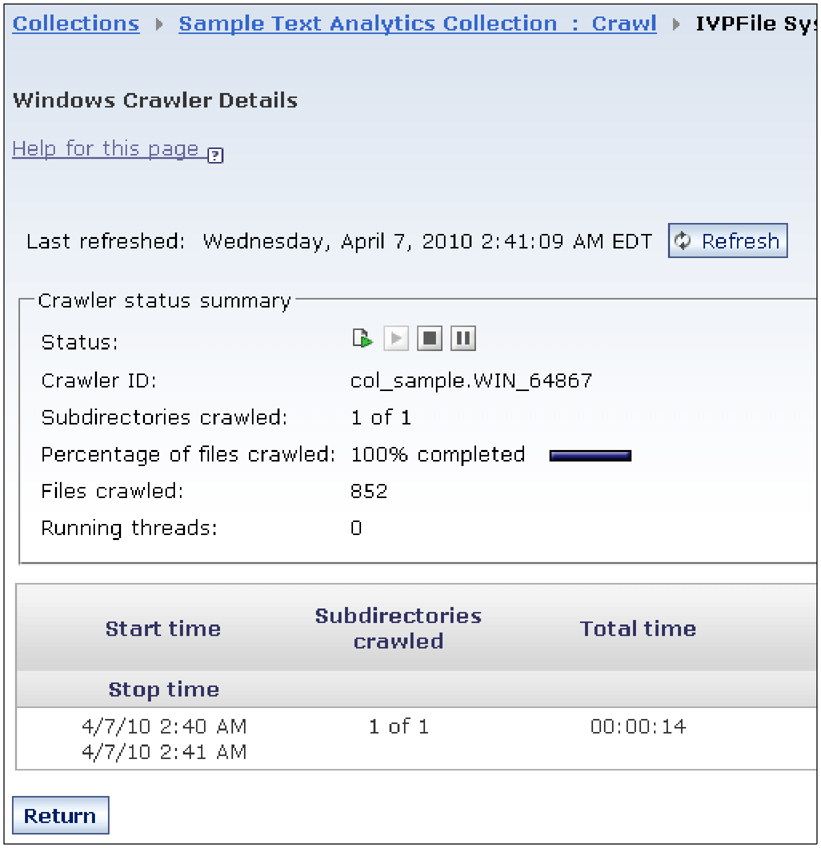
To check the status of the crawler by using the administration console, from the Collections view, click **Monitor** and then click the **Details** icon (Figure 4-71).



Crawler name	Type	Details	Status
IVPFile System Windows file system	Windows file system	🗎	▶ ■ ❚❚

Figure 4-71 Crawler in monitor mode

The status of the crawler is shown in the Windows Crawler Details window (Figure 4-72.) The window also shows the crawl start and stop time, the number of files and subdirectories crawled, and the percentage of data crawled.



Collections ▸ Sample Text Analytics Collection : Crawl ▸ IVPFile Sy

Windows Crawler Details

Help for this page

Last refreshed: Wednesday, April 7, 2010 2:41:09 AM EDT   Refresh

Crawler status summary

Status:	▶ ■ ❚❚
Crawler ID:	col_sample.WIN_64867
Subdirectories crawled:	1 of 1
Percentage of files crawled:	100% completed
Files crawled:	852
Running threads:	0

Start time Stop time	Subdirectories crawled	Total time
4/7/10 2:40 AM 4/7/10 2:41 AM	1 of 1	00:00:14

Figure 4-72 Status of the crawler component

Creating a custom crawler plug-in

Content Analytics provides an option to configure a custom crawler plug-in. A crawler plug-in is a Java program that is invoked for each document that is crawled. With this flexible feature, you can customize or change the behavior of a crawler.

For example, you might want to crawl a discussion on social network web pages. When a default web crawler provided by Content Analytics crawls the pages, it creates one document per web page. That web page might contain messages and replies from multiple users. In reality, the resulting crawled web page contains content from multiple sources. If you want to create one crawled content per source (or per user), you can create a custom plug-in that crawls the web pages. This plug-in also separates the crawled page into multiple documents (one document per user comment) and writes the documents to a file system. You can then use the default file system crawler to crawl the data into your system.

For more information about creating and configuring a custom crawler plug-in, go to the IBM Content Analytics Information Center available at the following address, and search on *crawler plug-ins*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

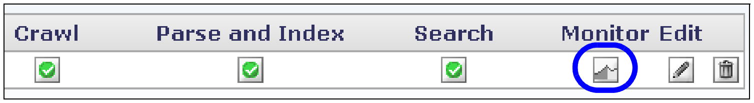
4.3.4 Building an index in the text analytics collection

The parse and index component is responsible for processing the data and building an index. By default, when a collection is created, the parse and index component is automatically started. The parse and index process continuously creates an index as long as the process is running and the crawler component is crawling the data. However, configuration changes to the collection after the index has been created requires restarting the parse and index component and rebuilding the collection. This section provides the steps to start and stop the parse and index component, redeploy the resources, and rebuild the full index.

Starting the parse and index component

To start the parse and index component, follow these steps:

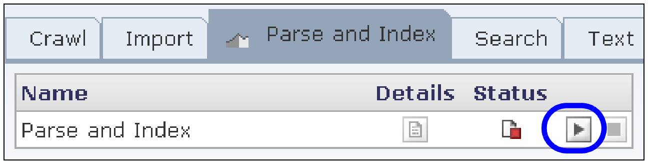
1. From the administration console, go to the **Collections** view and click the **Monitor** icon (Figure 4-73).



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓	📈	✎ 🗑

Figure 4-73 Collections view with editing and monitoring options

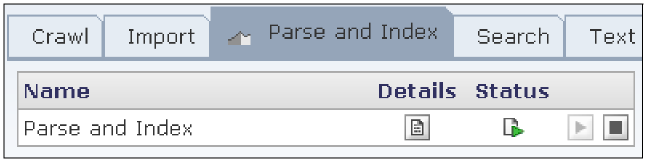
2. On the **Parse and Index** tab, click the **Start** icon (Figure 4-74) to start the component.



Name	Details	Status
Parse and Index		

Figure 4-74 Start icon on the Parse and Index tab

Figure 4-75 shows the status when the parse and index component is started.



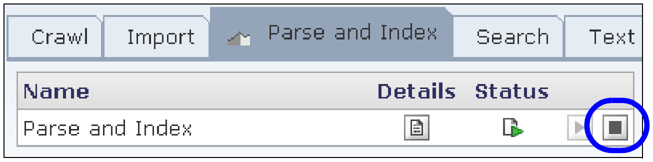
Name	Details	Status
Parse and Index		

Figure 4-75 Parse and Index in start mode

Stopping the parse and index component

To stop the parse and index component, follow these steps:

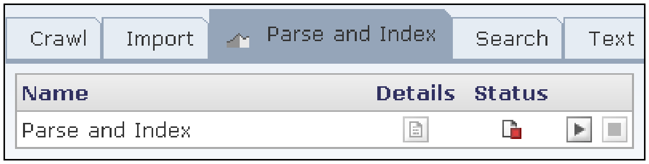
1. From the administration console, go to the Collections view and click the **Monitor** icon.
2. Select the **Parse and Index** tab and click the **Stop** icon (Figure 4-76) to stop the parse and index component.



Name	Details	Status
Parse and Index		

Figure 4-76 Stop icon on the Parse and Index tab

Figure 4-77 shows the status when the parse and index component is stopped.



Name	Details	Status
Parse and Index		

Figure 4-77 Parse and Index in stop mode

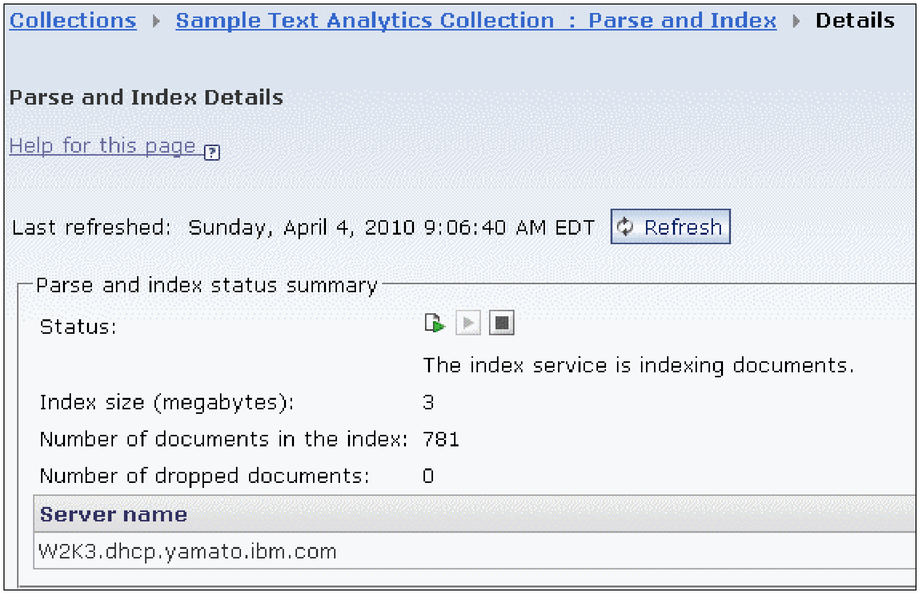
Checking the index and parse component status

The administration console provides the status of the index and parse component. The parse and index component has one of two statuses:

- The index service is indexing the document.
- The index service is waiting for crawled documents.

The index service is indexing the document

The "`The index service is indexing the document`" status typically means that the component is processing the document as shown in Figure 4-78. The Number of dropped documents field refers to the number of documents that the component cannot process for some reason.



Collections ▸ Sample Text Analytics Collection : Parse and Index ▸ Details

Parse and Index Details

Help for this page

Last refreshed: Sunday, April 4, 2010 9:06:40 AM EDT

Parse and index status summary

Status: The index service is indexing documents.

Index size (megabytes): 3

Number of documents in the index: 781

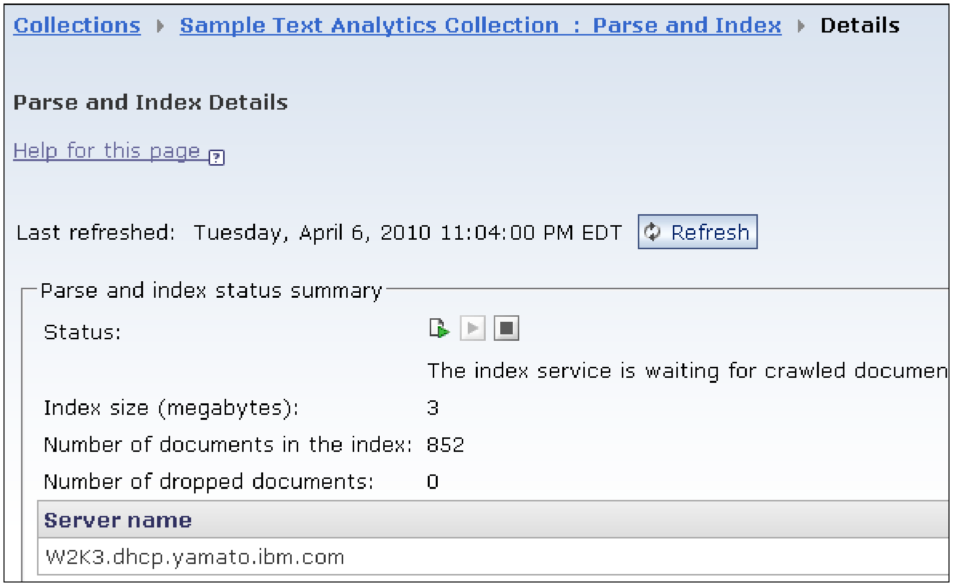
Number of dropped documents: 0

Server name
W2K3.dhcp.yamato.ibm.com

Figure 4-78 Parse and Index in a running state

The index service is waiting for crawled documents

As shown in Figure 4-79, the "The index service is waiting for crawled documents" status means that the parsing and indexing component has processed all the documents and is idle now. This status typically happens when all the documents have been processed.


Collections ▸ Sample Text Analytics Collection : Parse and Index ▸ Details

Parse and Index Details

Last refreshed: Tuesday, April 6, 2010 11:04:00 PM EDT

Parse and index status summary

Status:

The index service is waiting for crawled documen

Index size (megabytes): 3

Number of documents in the index: 852

Number of dropped documents: 0

Server name
W2K3.dhcp.yamato.ibm.com

Figure 4-79 Parse and Index in an idle state

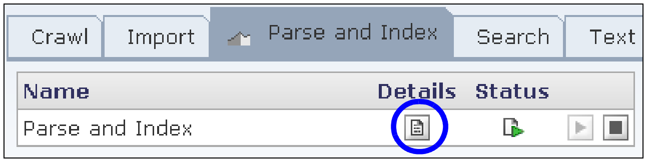
To verify whether all documents are processed, compare the number of documents crawled with the number of documents in the index. If these two numbers are the same, the document processing is complete. To find the number of documents crawled, see "Checking the crawler component status" on page 131.

Resource deployment

Typically, when you have a collection with data already crawled and analyzed, changes to configuration require restarting the system or components, which is time consuming. If you do not have any documents processed, you can use the resource deployment option to make changes to a facet tree, user dictionaries, or custom text analysis rules effective without rebuilding the index. For example, when we create facets and facet mappings in "Creating facets and mapping search fields to facets" on page 106, we can use resource deployment instead of rebuilding the index to make changes effective.

To redeploy resource, follow these steps:

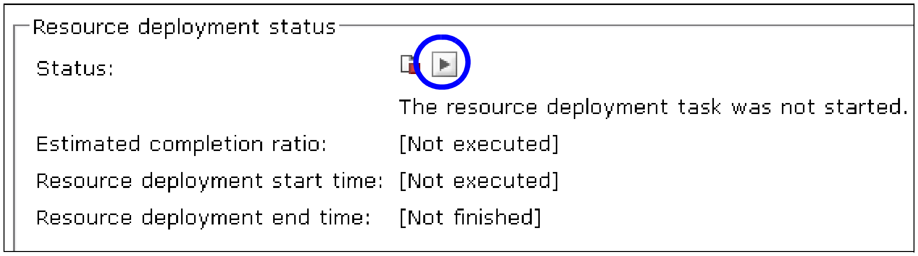
1. From the administration console, go to the Collections view and click the **Monitor** icon.
2. Select the **Parse and Index** tab.
3. Click the **Details** icon (Figure 4-80) to stop the parse and index component.



Name	Details	Status
Parse and Index		

Figure 4-80 Parse and Index in start mode

4. In the Resource deployment status section (Figure 4-81), click the **Start** button.


Resource deployment status

Status:

The resource deployment task was not started.

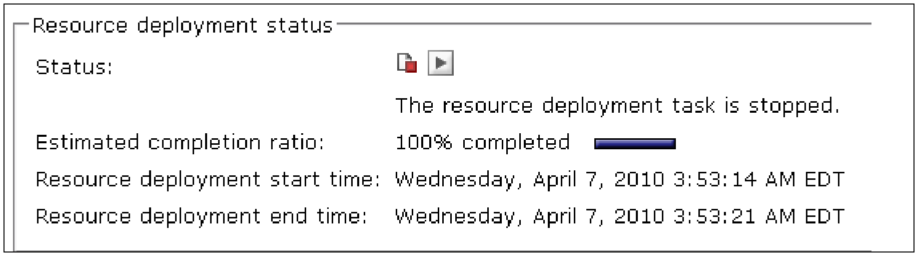
Estimated completion ratio: [Not executed]

Resource deployment start time: [Not executed]

Resource deployment end time: [Not finished]

Figure 4-81 Resource deployment status

5. Wait for the task to complete (Figure 4-82).


Resource deployment status

Status:

The resource deployment task is stopped.

Estimated completion ratio: 100% completed

Resource deployment start time: Wednesday, April 7, 2010 3:53:14 AM EDT

Resource deployment end time: Wednesday, April 7, 2010 3:53:21 AM EDT

Figure 4-82 Resource deployment status showing the task is complete

Rebuilding the full index

When you iterate through designing a text analytics collection with a subset of data, the document processor component might require configuration changes such as defining new search fields. This type of change often affects how the index is built. Therefore, you are required to rebuild a full index.

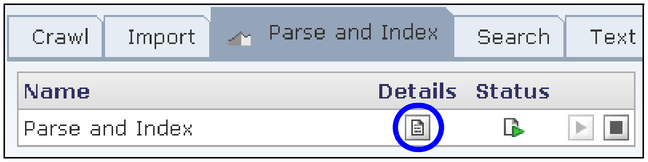
When you change any of the following parts of the configuration, you must rebuild the index:

- Parsing options
- Text processing options or the document processing pipeline
- Additions or modifications to the facets or search fields
- Native field to search field mappings
- User dictionaries and text analysis rules

Updating native fields to search fields mappings: When you update the native field to search field mapping, especially when the XML to search fields mapping is changed, the XML documents must be recrawled because redeploying the resources does not pick up changes made to XML mappings.

To rebuild a full index using the administration console, follow these steps:

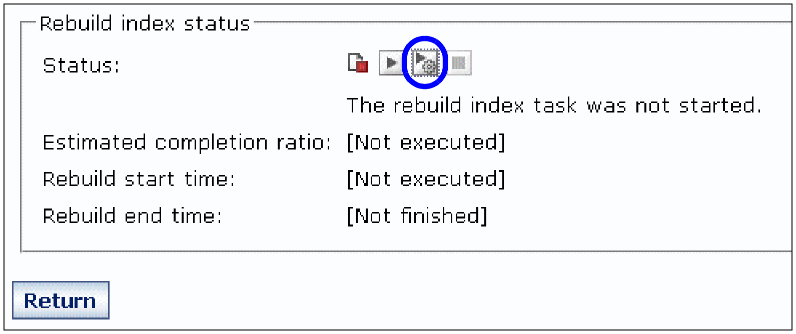
1. From the administration console, go to the **Collections** view and click the **Monitor** icon.
2. On the **Parse and Index** tab, click the **Details** icon (Figure 4-83).



Name	Details	Status
Parse and Index		

Figure 4-83 Parse and Index tab

3. Click the **Restart a full index build** icon (Figure 4-84).



Rebuild index status

Status: The rebuild index task was not started.

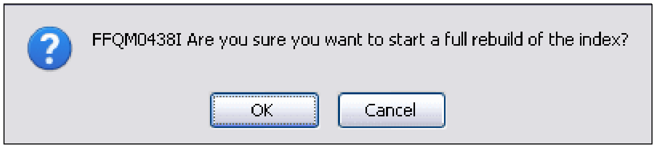
Estimated completion ratio: [Not executed]

Rebuild start time: [Not executed]

Rebuild end time: [Not finished]

Figure 4-84 Building a full index

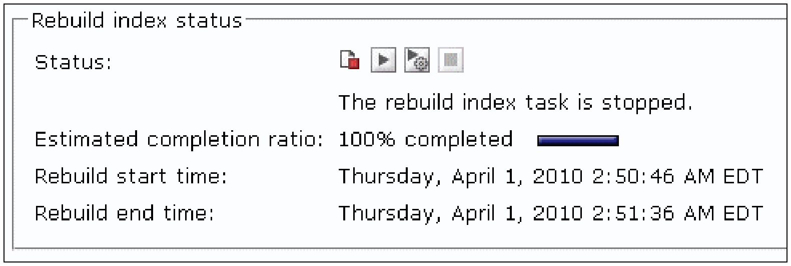
4. In the confirmation message window (Figure 4-85) that opens, click **OK**.



FFQM0438I Are you sure you want to start a full rebuild of the index?

Figure 4-85 Full rebuild confirmation message window

5. Wait for the rebuild index to complete as shown in Figure 4-86.



Rebuild index status

Status: The rebuild index task is stopped.

Estimated completion ratio: 100% completed

Rebuild start time: Thursday, April 1, 2010 2:50:46 AM EDT

Rebuild end time: Thursday, April 1, 2010 2:51:36 AM EDT

Figure 4-86 Rebuilding index completed

4.4 Verifying that the collection is available

This section explains how to confirm that the created text analytics collection is working as expected with the text miner application.

4.4.1 Starting the search server for the text analytics collection

After the index of the text analytics collection is ready, start the search session for the collection from the administration console. By default, the search session is started when you create a text analytics collection. However, you might stop the search session when you configure the text analytics collection. Make sure that the search session is up and running before you verify the collection with the text miner application.

4.4.2 Accessing the collection with the text miner application

After the index is ready for your analysis, you can verify whether the text analytics collection is available for your analysis by accessing the text miner application. Make sure that at least one text analytics collection is available for search before you launch the text miner application.

You can launch the text miner application from the following URL:

```
http://<Content Analytics Server Host Name:Port Number>/analytics/
```

For more information about how to work with the text miner application, see Chapter 5, “Text miner application: Basic features” on page 143.

4.5 Deploying the configuration

When you work with Content Analytics, test your configuration with a small set of data on the test environment. When you feel comfortable with your setup, you can move your configuration from the test environment to the production environment. This section explains how to move your existing configuration to the new system.

4.5.1 Using the esadmin export and import commands

The convenient way to move the collection configuration is to use the **esadmin export** and **import** commands.

With the **esadmin export** command, you can export a collection configuration to a compressed (.zip) file. You can then import the collection configuration from the compressed file to another system by using the **esadmin import** command. The file size of the compressed file created by the **esadmin export** command is relatively small and easy to copy between the systems because the **esadmin export** command does not export the index data.

When you run the **esadmin export** command without specifying the output directory, the compressed file is created under the ES_NODE_ROOT/dump directory as the *collectionID*_export_*yyyymmdd_mmhhssz*.zip file. In this file name, *collectionID* is the correction ID that you specify with the **-cid** option, *yyyymmdd_mmhhss* represents the date and time when the export is performed, and *z* is the time zone offset from GMT.

After the compressed file is created, you copy the file to the target system and specify the file with the **-fname** option of the **esadmin import** command.

Deleting the index created by the text analytics collection: You can use the same **esadmin export** and **esadmin import** commands to delete the index created in the text analytics collection on the same system:

1. Export the collection configuration by using the **esadmin export** command.
2. Remove the collection from the administration console.
3. Import the collection configuration by using the **esadmin import** command with the exported file. Specify the same collection name or same collection ID.

   If you want to overwrite the existing collection with the same collection name instead of removing the collection, use the **-force** option with the **esadmin import** command. However, when you use the **-force** option, you must be aware of several important items. For more information, go to the IBM Content Analytics Information Center at the following address, and search on *exporting and importing collection configurations*:

   http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp
4. To create an index in a newly created collection, run all the processes (crawling, parsing, and indexing) again after you import the collection configuration.

4.5.2 Using the esbackup and esrestore commands

If you want to back up a collection configuration file and the entire index, you can use the **esbackup** and **esrestore** commands. Saving the backup image consumes a lot of disk space. The **esbackup** and **esrestore** commands back up and restore everything, including the host name in the configuration file.

These commands are used for backup and restore purposes such as for disaster recovery. They are not used for moving configuration files to another system. Saving the backup image consumes a lot of disk space.

4.5.3 Usage guidelines for these commands

The **esadmin export** and **esadmin import** commands are convenient in the following situations:

- When you move the configuration files to another machine
- When you reconfigure the collection again on the same machine

You can run these commands even when you are performing the analysis. Consider the following tips for using the commands:

- Only the collection-level configuration files are exported. That is, if you use the custom annotator (processing engine archive (PEAR) file) in the collection, the **esadmin export** command does not archive the PEAR file in the compressed file. You must configure to load the PEAR file on the target system from the administration console by yourself. You must also keep the name the same as the system name from which you export the configuration.
- You must use the *same* version of Content Analytics for both the source system (where you export the configuration from) and the target system (where you import the configuration to).
- You must configure the crawler properly on the target system. Importing the configuration alone does not guarantee that the crawler will work as expected with the target data source on the new system.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *exporting and importing collection configurations*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

In some situations, you can use the **esbackup** and **esrestore** command utilities. When using these utilities, keep in mind the following considerations:

- All sessions are stopped when the backup and restore script is running. Therefore, you cannot analyze the data in the text analytics collection while the backup and restore is running.
- You need at least the same (or much larger) disk space to back up the data and move the data to the new system to restore it. Usually the backup and restore task takes a lot of time and disk space.
- You must use the *same* disk layout. For example, if the data is in the `E:\data` path on the original system, the target system must have the same disk layout (`E:\data`) so that the data is restored in the same directory.
- You must use the *same* version of Content Analytics for both the source system (where you take the backup) and the target system (where you restore the backup).

- You must configure the crawler correctly on the target system where the backup is restored. Especially when you crawl the content from the local file system, you must have the same target directory to crawl.

The configuration is preserved from the source system. It might be better to have the same configuration (such as the installation directory, disk layout, and crawler configuration) as much as possible so that you can easily manage the target system.

For more information, go to the IBM Content Analytics Information Center at the following addresses and search on the following topics:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

- “Backing up the system”
- “Restoring the system”

5

Text miner application: Basic features

Chapter 4, “Installing and configuring IBM Content Analytics” on page 71, explains how to build a text analytics collection and configure it for analysis. This chapter focuses on the basic features of the text miner application, which is a web-based application that helps you to discover actionable insight from your textual data. This chapter provides details about the application, focusing on the search and discovery features. For information about the text miner application views, which are also part of the basic features, see Chapter 6, “Text miner application: Views” on page 217.

This chapter includes the following sections:

- Overview of the text miner application
- Search and discovery features
- Common view features
- Document flagging

If you are familiar with the user interface of the text miner application, including the search and discovery features, and all its views, proceed with Chapter 7, “Performing content analysis” on page 279.

5.1 Overview of the text miner application

This section provides an overview of the text miner application by covering the basic application window layout and functionality. The sections that follow go into greater detail about the various forms of analysis that you can perform by using the sample data set packaged with IBM Content Analytics.

> **Sample Text Analytics Collection:** Most of the examples in this chapter refer to the Sample Text Analytics Collection that is created when you select **Text Analytics Tutorial** in the First Steps program. Alternatively, you can build the text analytics collection with the same data and configuration as instructed in Chapter 4, “Installing and configuring IBM Content Analytics” on page 71. You must build this collection (if you have not already built one), so that you can follow along using your installed version of Content Analytics.

5.1.1 Accessing the text miner application

The text miner application is a Java 2, Enterprise Edition (J2EE), web-based application that is automatically deployed onto the search server when you install Content Analytics. Before you access the text miner application, ensure that at least one text analytics collection is available for search by using the administration console, as explained in Chapter 4, “Installing and configuring IBM Content Analytics” on page 71.

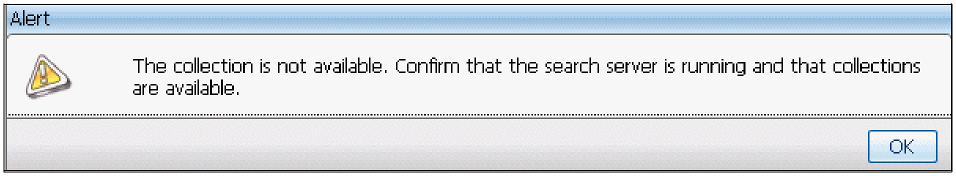
You can access the text miner application at the following address:

```
http://<Content Analytics Server Host Name:Port Number>/analytics/
```

Content Analytics Server Host name is the server on which you install Content Analytics. *Port Number* is the number that you specified during installation. By default, the port number is 8393. If you install Content Analytics with a multiserver installation, you can access the text miner application that is installed on each search server.

If no collection search run times are started, you see the following Alert message, which is also shown in Figure 5-1:

```
"The Collection is not available. Confirm that the search server is
running and that collections are available."
```



Alert

The collection is not available. Confirm that the search server is running and that collections are available.

Figure 5-1 Alert message that is displayed if no text analytics collection is available

When you install Content Analytics on Windows platform, you can also access the text miner application from the Start menu on the Content Analytics server by selecting **Start** → **All Programs** → **IBM Content Analytic** → **Text Miner Application**.

5.1.2 Application window layout and functional overview

When you open the text miner application with a browser, you see a window similar to the one in Figure 5-2 on page 146. The text miner application has the following areas:

- Application toolbar

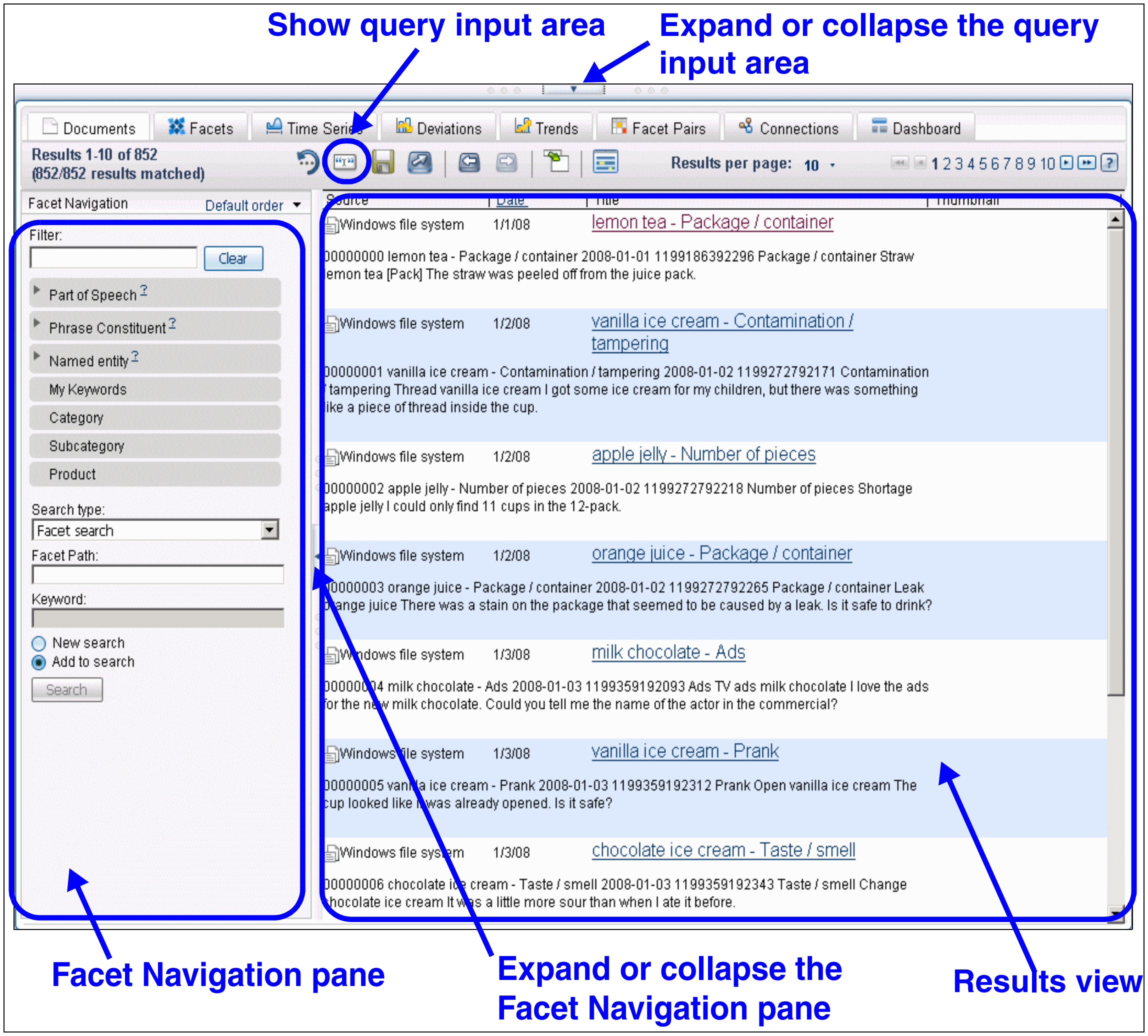
  This toolbar is at the top of the window (not shown in Figure 5-2 on page 146). You use it to select the collection to analyze and set preferences. You can also use it to obtain help at anytime (Figure 5-3 on page 147).

- Query search text field and controls

  The search box and controls are located beneath the application toolbar. You use them to build and add to your search expression. Content Analytics shows only those documents that match your current query conditions. This area is hidden by default. To reveal the search text field and controls, you click the **Show query input area** icon (circled in Figure 5-2 on page 146).

- Facet Navigation pane

  The Facet Navigation pane is on the left side of the application window. You use this area to filter the results based on selected facets and keywords. This pane is shown by default. To collapse it, you click the **Collapse this area** button between the Facet Navigation and Results view panes (Figure 5-2 on page 146).



Source	Date	Title
Windows file system	1/1/08	lemon tea - Package / container

00000000 lemon tea - Package / container 2008-01-01 1199186392296 Package / container Straw lemon tea [Pack] The straw was peeled off from the juice pack.

Source	Date	Title
Windows file system	1/2/08	vanilla ice cream - Contamination / tampering

00000001 vanilla ice cream - Contamination / tampering 2008-01-02 1199272792171 Contamination / tampering Thread vanilla ice cream I got some ice cream for my children, but there was something like a piece of thread inside the cup.

Source	Date	Title
Windows file system	1/2/08	apple jelly - Number of pieces

00000002 apple jelly - Number of pieces 2008-01-02 1199272792218 Number of pieces Shortage apple jelly I could only find 11 cups in the 12-pack.

Source	Date	Title
Windows file system	1/2/08	orange juice - Package / container

00000003 orange juice - Package / container 2008-01-02 1199272792265 Package / container Leak orange juice There was a stain on the package that seemed to be caused by a leak. Is it safe to drink?

Source	Date	Title
Windows file system	1/3/08	milk chocolate - Ads

00000004 milk chocolate - Ads 2008-01-03 1199359192093 Ads TV ads milk chocolate I love the ads for the new milk chocolate. Could you tell me the name of the actor in the commercial?

Source	Date	Title
Windows file system	1/3/08	vanilla ice cream - Prank

00000005 vanilla ice cream - Prank 2008-01-03 1199359192312 Prank Open vanilla ice cream The cup looked like it was already opened. Is it safe?

Source	Date	Title
Windows file system	1/3/08	chocolate ice cream - Taste / smell

00000006 chocolate ice cream - Taste / smell 2008-01-03 1199359192343 Taste / smell Change chocolate ice cream It was a little more sour than when I ate it before.

Figure 5-2 Text miner application window layout

- Result view

  The Results view is located in the center and to the right of the Facet Navigation pane. This area shows the results that match your current search and facet selections. The view changes depending on which of the following view tabs you select:

  – Documents view
  – Facets view
  – Time Series view
  – Trends view

- Deviations view
- Facet Pairs view
- Connections view
- Dashboard view

Dashboard view: The Dashboard view requires additional configuration. See 6.9, “Dashboard view” on page 262 for more details.

See Chapter 6, “Text miner application: Views” on page 217, for more details about each of these views.

Application toolbar

The application toolbar is at the top of the text miner application window. This toolbar shows the name of the current text analytics collection that is under analysis. You can easily switch to another collection for analysis by clicking the **change** link (Figure 5-3).

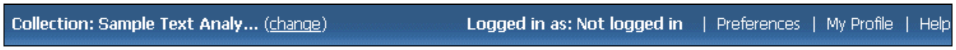

Collection: Sample Text Analy... (change)
Logged in as: Not logged in
| Preferences
| My Profile
| Help

Figure 5-3 Application toolbar

The Logged in as field shows the name of the currently logged in user. If security is not enabled, the “`not logged in`” message is displayed.

More information: See Appendix A, “Security in IBM Content Analytics” on page 633, for details about security in Content Analytics.

When you click the **Preferences** link, another window is displayed in which you can set various options for Search, Results, Result Columns, and each of the views.

When you click the **My Profile** link, you see a table of security credentials to use when accessing secured data sources. If security is enabled for the application server and your text analytics collection was created with security enabled, you need to specify access credentials (user IDs and passwords) for each data source that requires secured access.

At anytime during the operation of the text miner application, you can click the **Help** link for assistance, which opens a new browser window. When accessing Help, the IBM Content Analytics Information Center must be open on the Content Analytics server that is running.

Assumption: This chapter was written with the understanding that security is not enabled.

Information center: By default, IBM Content Analytics Information Center on the Content Analytics server starts when you start the Content Analytics server by using the `esadmin system startall` command.

Search box and controls

The search box under the application toolbar is used to find and filter documents based on your queries. Content Analytics contains a fully functioned search engine that is both scalable and fast so that you can explore your data by using conventional search methods.

By default, the search box is hidden. After you click the **Expand query input area** button, you see the search box (Figure 5-4). You can modify the query that is used for your analysis from this field.

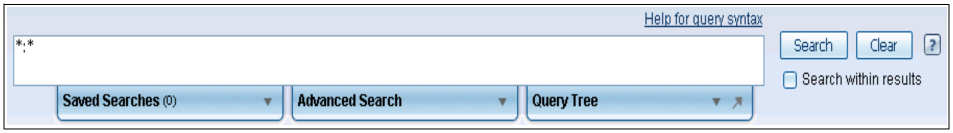


Figure 5-4 Search fields and controls

The current query is displayed in the search box. Each addition to your search condition is appended to the query. At anytime, you can save your current query state by clicking the **Saved Searches** tab. After giving your query a name and saving it, you can return to that state by clicking the **Saved Searches** tab and selecting that name.

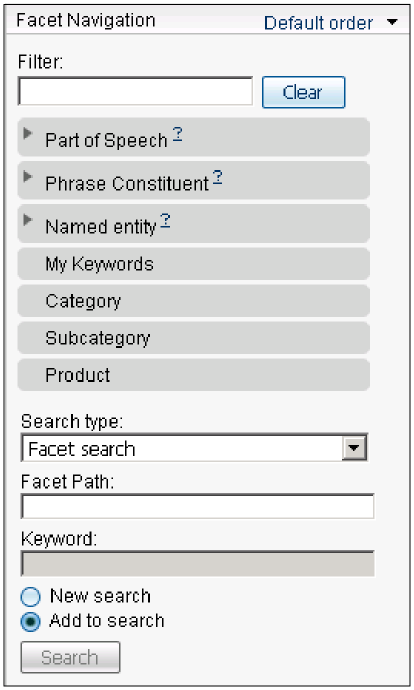
You can use the **Advanced Search** tab to assist in the formulation of your search expression. By using the **Advanced Search** tab, you can specify complex search options, such as an exact phrase search or a search on date ranges, without knowing the Content Analytics query syntax.

A graphical layout of your query is displayed when you click the **Query Tree** tab.

For more information about these tabs, see 5.6, “Common view features” on page 203.

Facet Navigation and search filter

The Facet Navigation pane is always present and displayed on the left when you access the text miner application. In this pane, you see the facet tree that you defined for the text analytics collection in the administration console. The facet tree includes a list of predefined facets, such as Part of Speech and Phrase Constituent, as shown in Figure 5-5.



Search type: Facet search

○ New search
● Add to search

Figure 5-5 Facet Navigation pane and search filter

You can click a facet in the Facet Navigation pane to select it, at which point the selected facet is highlighted in blue. Also several controls are available to assist you in locating a specific facet. These controls are useful when the number of facets is large. You can filter which facets to display by the name of the facet or by the actual value of a particular facet.

Text miner application views

You can select from six result views (Figure 5-6), depending on your analysis goals, and interact with each view to drill down further into the data. For a detailed description of each view, see Chapter 6, “Text miner application: Views” on page 217.

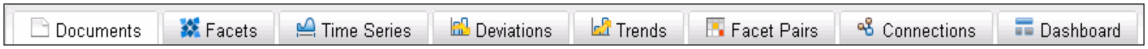


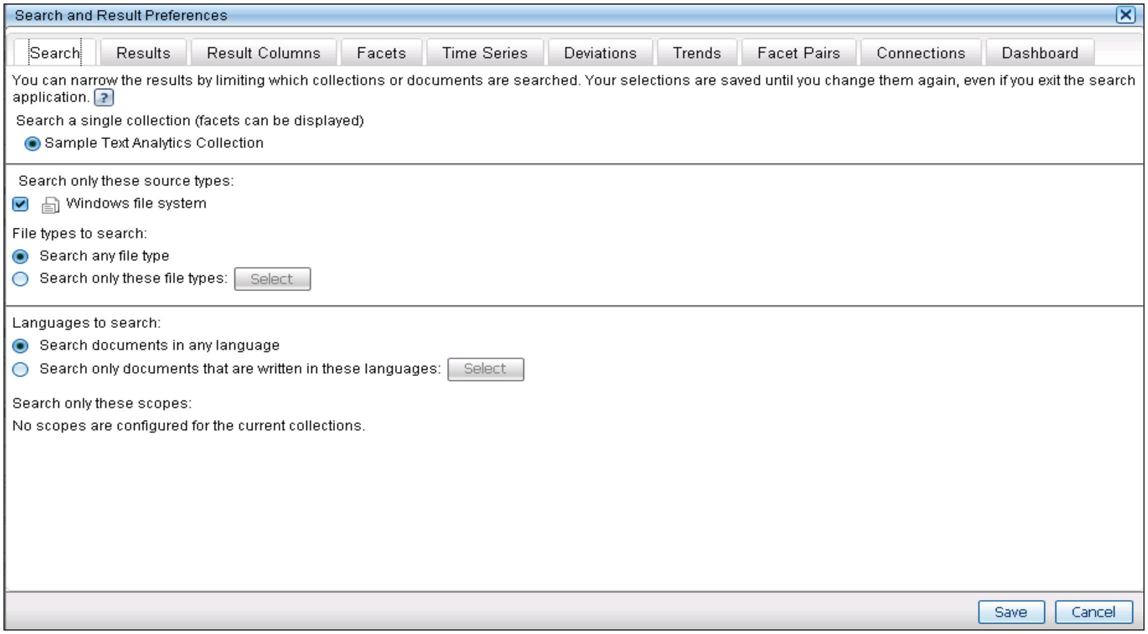
Figure 5-6 Tabs for each view, with the Time Series view selected

5.1.3 Selecting a collection for analysis

To start analyzing your data with the text miner application, you must select a collection from the application toolbar by clicking the **Change** link. You can analyze only one collection at a time with the text miner application.

5.1.4 Changing the default behavior by using preferences

You can change the default behavior of the text miner application by using the Preferences window that is accessible from the application toolbar. The changes made by using the Preferences window remain in effect for the duration of your browser session. Otherwise, if global security is enabled, they are persistently saved in your profile. After you click the **Preferences** link, the Search and Result Preferences window (Figure 5-7) opens.



You can narrow the results by limiting which collections or documents are searched. Your selections are saved until you change them again, even if you exit the search application. ?

Search a single collection (facets can be displayed)

● Sample Text Analytics Collection

Search only these source types:

☑ Windows file system

File types to search:

● Search any file type

○ Search only these file types: Select

Languages to search:

● Search documents in any language

○ Search only documents that are written in these languages: Select

Search only these scopes:

No scopes are configured for the current collections.

Figure 5-7 Search and Result Preferences window

This section highlights the options that are available in the Search and Result Preferences window, in particular, the Search, Results, and Result Columns tabs. These three tabs affect the results that are displayed in the Documents view and help you to review the details of the documents that are selected.

You can also set the preferences for other views, such as the Facets view, Time Series view, Deviations view, Trends view, Facet Pairs view, Connections view, or Dashboard view from the Preferences window.

Search tab

From the **Search** tab, you must select a text analytics collection to analyze. You also select the following properties on this tab:

- The source type that you want to search.
- The file types to search. By default, **Search any file type** is selected. You can also specify another file type here.
- The language to search. By default, **Search documents in any language** is selected. You can specify the language to search if you focus on specific documents written in a specific language.
- Search only the specific scopes. This option is available only when you define a search scope.

Results tab

In the **Results** tab, you control how the results are displayed in the Documents view. The change takes effect right after it is saved. The properties are useful when you perform the search from the Search box field and are important when you filter your data with a query.

You can select the following properties from this tab:

- Number of results per page. By default, 10 results per page are displayed. If you click the **count up** or **count down** button, the value increases by 5. You can also specify a discrete number in the box.
- Query Language. By default, the same value is set that you set in the Language to use field on the **Search** tab.
- Query mode. You can select the linguistic parsing method for the query that you input. By default, this option is set to **No preferences**. For the difference of each mode, see the help information.
- Sort by. By default, **Relevance** is selected. All search fields that are specified as sortable in the administration console are displayed in the selectable list.
- Sort order. After you select a search field in the Sort by field, you can select the sort order as either descending or ascending.
- Number of type-ahead suggestions. By default, 10 suggestions are specified from the type-ahead feature.

- Type-ahead mode. By default, **Suggest matches from queries, then matches from the index** is selected. You can disable the type-ahead suggestion or select another suggestion order.
- Summary length. By default, the length of the summary is set as a medium value, but you can adjust the length of the summary by a grade of 5.
- Include quick links. By default, **No** is selected.
- Suggest spelling corrections. By default, **Yes** is selected.
- Show a file type filter. By default, **No** is selected.
- Show icons to list documents similar to a selected document. By default, **No** is selected.
- Similarity. If you enable document similarity, set the Similarity field value to define how similar documents must be in order to be identified as being similar to one another.
- Show categories when you preview documents. By default, **Yes** is selected.
- Show category rules when preview documents. By default, **No** is selected.

Result Columns tab

On the **Result Columns** tab, you select which columns to display in the Documents view and the order of the columns. You can select or clear the column name and change the column order.

When you configure the system to use document flags or query builder, customizing the column order is helpful. The document flags functionality adds a Flags column to the Documents view, and the query builder adds an Actions column. The order of these columns can be modified.

5.2 Search and discovery features

The text miner application is a business intelligence tool that is used to gain valuable insight from your data. Text mining itself is an exploratory task with basic goals set by you. The goals that you define serve as a guide for the strategies that you employ to explore your data.

At the core of the Content Analytics product is a search engine that is scalable, fast, and helps you meet your business intelligence goals. The search engine is based on the open source Lucene indexer that is enhanced with extensions made by IBM. It is uniquely engineered to support rapid search and discovery of your data.

The exploration of your data is one of the iterative steps taken through the many dimensions of your data. Content Analytics guides you through the exploration of these various dimensions, alerting you to any interesting patterns or trends that warrant further investigation.

When you initially start the text miner application and choose a text analytics collection for analysis, your search query (which is displayed in the search box) is set to the wildcard expression `*:*`. This search expression results in all documents in the collection to be matched and thus equates to all documents in the collection being analyzed. From this starting point, you can browse through the documents by using the Documents view, or you can see how all of the documents are distributed over time by using the Time Series view. You can also start to explore the various facets of your data by using the Facets view.

Eventually you can narrow the scope of your analysis to a more focused subset of documents. Narrowing down the documents is achieved by using facet selection, search expressions, or a combination of both.

This section provides details about the following search and discovery features:

- Limiting the scope of your analysis using facets
- Limiting the scope of your analysis using search operators
- Limiting the scope of your analysis using dates
- Query syntax
- Type ahead
- Saved searches
- Advanced search

For other search and discovery topics, see the following sections:

- 5.3, “Query tree” on page 163
- 5.4, “Query builder” on page 180
- 5.5, “Rule-based categories with a query” on page 193

5.2.1 Limiting the scope of your analysis using facets

In the Facet Navigation pane, you can select a particular facet for viewing. Typically you use the Facets view to see the most frequently occurring values found in your data for a particular facet. For example, you can have a facet labeled “countries.” When you select the Facets view, the system lists the top 500 most frequently occurring country names mentioned in your textual data. The 500 view limit is the default, which you can change by using the Preferences window as previously mentioned.

At any time, you can select one or more keywords and add them to your query to limit the scope of your analysis. In our country facet example, if we select the keywords Mexico and Canada and add them to our query, only those documents that mention Mexico or Canada are used in analysis by the text miner application.

You can change the way the keywords are added to your current query expression by using the Boolean logic: AND, OR, and AND NOT. In each of the text miner result views (except for the Documents view), you see the search Boolean operators icons: AND, OR, and AND NOT. Each icon gives a visual representation of a Venn diagram depicting the type of Boolean operation to be performed as described in Table 5-1.

Table 5-1 Boolean operators and their description

Icon	Description
AND	This operator generates the appropriate search syntax for the keywords that you select and appends it to your query with an AND condition. The result is documents that match your query condition *and* contain any of the selected keywords.
OR	This operator generates the appropriate search syntax for the keywords that you select and appends it to your query with an OR condition. The result is documents that match your query condition *or* the presence of any of the selected keywords.
AND NOT	This operator generates the appropriate search syntax for the keywords that you select and appends it to your query with an AND NOT condition. This method is a convenient way to exclude certain documents from your analysis.

For examples of using these search operators, see 5.3.2, “Understanding the query tree” on page 165.

5.2.2 Limiting the scope of your analysis using search operators

Selecting facets is one way to constrain the scope of your analysis to only those documents that match a given keyword. However, facets must be contrived and defined in advance when you are building your text analytics collection.

What if during your analysis with the text miner application, a facet is not defined for the type of constraint you want to filter on? In this case, you can use the powerful search features of Content Analytics to limit your analysis to only those documents that match your query. The filtering of documents based on your search expression is in addition to the facets that you have already selected.

Content Analytics provides a comprehensive query syntax with a robust set of search operators. You can use the Advanced Search function to assist you in your search expression. You can also use the Query Tree function to view the logical structure of your query as it grows more complex. When you turn on global security, at anytime you can also save the current state of your queries for future use during a browser session or persistently across a browser session. Saving queries is a convenient way for you to go back to the analysis that you previously did.

For more information about query syntax and query tree, see 5.2.4, “Query syntax” on page 156.

5.2.3 Limiting the scope of your analysis using dates

If your documents contain date fields, you can limit the scope of your analysis to documents that match specific dates or that fall between a given date range. The date range facet can be used to analyze the data within a given range.

You can specify date constraints by using the **Advanced Search** tab or by manually entering a parametric (date) search expression. An easier way to select specific dates or ranges is to use the Time Series view as explained in 6.4, “Time Series view” on page 226. In addition, when you configure a date range facet, you can use the data range facet to analyze the data within a given range.

A similar and important feature of Content Analytics is its ability to identify trends and patterns in your data that occur over time. In order for Content Analytics to identify trends and patterns, it must base its calculations on a certain date value that consistently occurs in each document of the text analytics collection.

With the date field, Content Analytics performs time-sensitive calculations and renders the various views including Time Series, Deviations, and Trends views. For more information about the various views, see Chapter 6, “Text miner application: Views” on page 217.

In addition, when your documents contain multiple date fields, you can use the various date fields for your analysis after you configure the date facet. By changing the date facet to use for analysis, you can analyze the same data from another aspect based on a different date field.

Configuring the date facet: To configure the date facet to contain more than the date field, see “Optional: Configuring the date facet” on page 113.

5.2.4 Query syntax

As your analysis progresses, you notice in the search box the precise query expression for the current set of documents being analyzed. Usually you do not need to consider the detailed query syntax itself. Other times you might find it useful to modify or add to the search expression manually as you become more comfortable with the query syntax. The query syntax supports many search operators. For example, you might want to use the following query syntax to help you narrow down the target document set and discover the data:

- Faceted Path search

  You can use the query with facet name and its path. For example, the query `[/"keyword$.product"/"apple juice"]` returns the documents in the Product facet with the keywords "`apple juice`."

  > **Facet path used in query:** Because Content Analytics uses an internal representation for the facet path, it might be difficult to construct the faceted path search yourself when you use the text miner application. However, you can confirm how the Faceted Path search is constructed when you add the facets with search operators in other views. You can see how the actual query keyword is displayed from the search field when you expand the query tree.

- Proximity search

  You can search a keyword within a specified number of keywords or within a sentence. For example, the query `(cream dirty) WITHIN 8` returns the documents that have those keywords within eight words of each other (and in any order).

  If you need to consider the word order, you can add `INORDER` at the end of the query, such as `(cream dirty) WITHIN 8 INORDER`. This way the query returns the documents that include the keyword within specified word gap in order.

  Alternatively, the query `(purchase ice cream) WITHIN SENTENCE` returns the documents that include all query keywords in the same sentence.

- Fuzzy search

  You can set the ambiguity with the ~ operator. For example, the query `apple juice~0.5` returns the documents that include `apple juice`, `apple juicer`, and so on.

- Wildcard search

  You can replace some part of the query keyword or phrase with wildcard characters, such as a question mark (?) or an asterisk (*). For example, the query `* juice` returns documents that include `apple juice`, `pine juice`, and so on. Also, the query `bot???` returns the documents that include `bottle`, `bottom`, and so on.

- Conceptual search

  If you configure IBM Classification Module integration or document clustering, you can perform the conceptual search. For more details, see 9.2.3, “Using a conceptual search for advanced content discovery” on page 364.

For details about the query syntax, go to the IBM Content Analytics Information Center at the following address, and search on *query syntax*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

> **Help for the query syntax:** If the information center is running on your Content Analytics server, you can access the query syntax help from the **Help for query syntax** link in the text miner application.

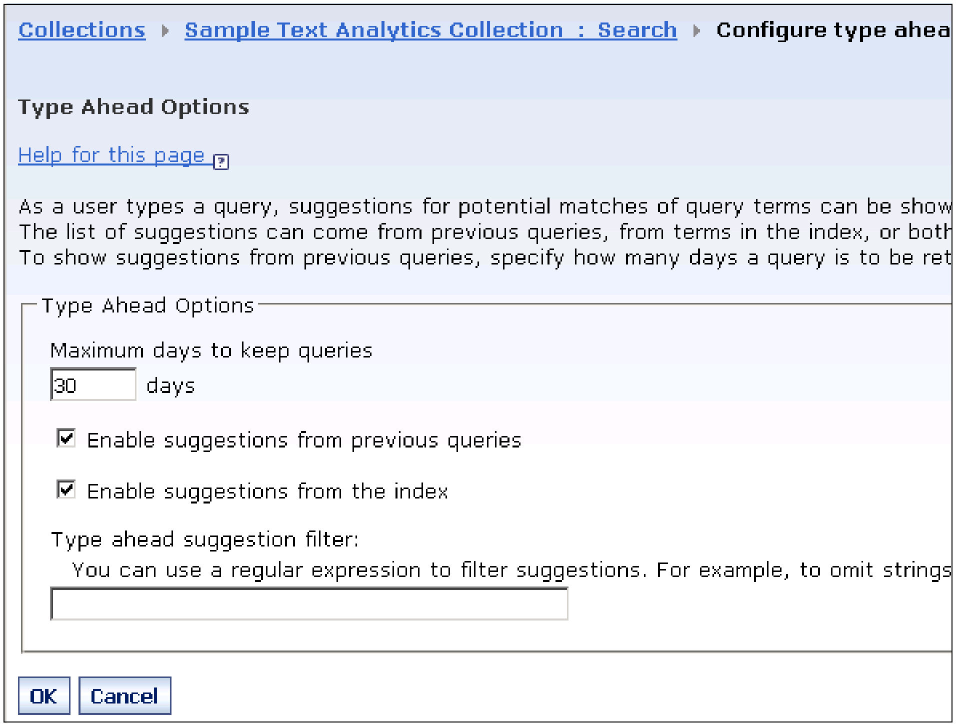
5.2.5 Type ahead

The type-ahead feature helps you to find the query keyword that you can use in your analysis. You can use the type-ahead feature when the text analytics collection is configured to enable the type-ahead feature. With the type-ahead feature, query suggestions based on the indexed terms or previous queries are shown as you type the query in the text miner application.

Configuring the type-ahead feature

The type-ahead feature is automatically enabled. You can configure where the type-ahead suggestions come from. By default, the query suggestions from both indexed terms and previous queries are built in the query index. You configure the type-ahead options from the administration console as shown in Figure 5-8 on page 158.

> **Type-ahead feature from previous queries:** To use the previous queries as suggestions for the type-ahead feature, the query log index must be enabled in the Advanced option when the text analytics collection is created, as explained in 4.3.2, “Creating a text analytics collection” on page 90. Otherwise, you can use the type-ahead suggestion from terms in the index only.



Collections ▸ Sample Text Analytics Collection : Search ▸ Configure type ahea

Type Ahead Options

Help for this page

As a user types a query, suggestions for potential matches of query terms can be show
The list of suggestions can come from previous queries, from terms in the index, or both
To show suggestions from previous queries, specify how many days a query is to be ret

Type Ahead Options

Maximum days to keep queries

30 days

☑ Enable suggestions from previous queries

☑ Enable suggestions from the index

Type ahead suggestion filter:

You can use a regular expression to filter suggestions. For example, to omit strings

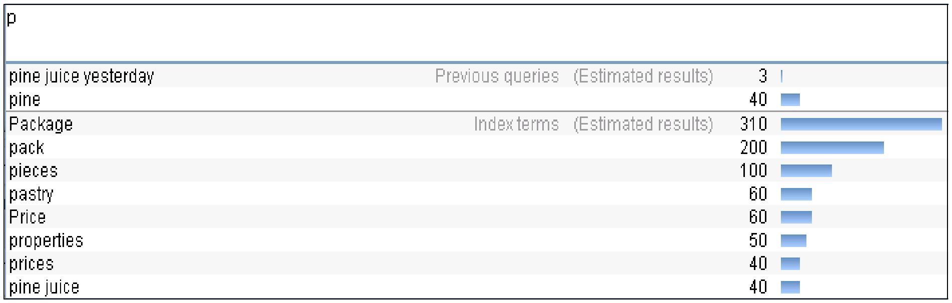
Figure 5-8 Configuring the type-ahead feature on the administration console

For details about the type-ahead feature configuration, go to the IBM Content Analytics Information Center at the following address, and search on *type ahead support for queries*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Using the type-ahead feature in the search box

When the type-ahead feature is configured, you see the query suggestion, as shown in Figure 5-9 on page 159. For example, when you type p in the query input area, the query suggestions are displayed based on your settings in Preferences. You can configure how many suggestions are displayed in the window and in which order the suggestions are displayed (suggestions from previous queries and then suggestions from index terms).

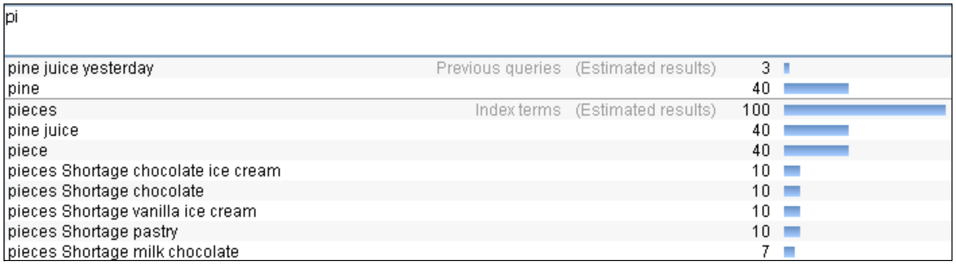


p

pine juice yesterday	Previous queries	(Estimated results)	3
pine			40
Package	Index terms	(Estimated results)	310
pack			200
pieces			100
pastry			60
Price			60
properties			50
prices			40
pine juice			40

Figure 5-9 Query suggestions when you type a character

When you type more than one character, such as `pi`, in the search query field, the query suggestions are narrowed to those words that match the query that you typed, as shown in Figure 5-10.



pi

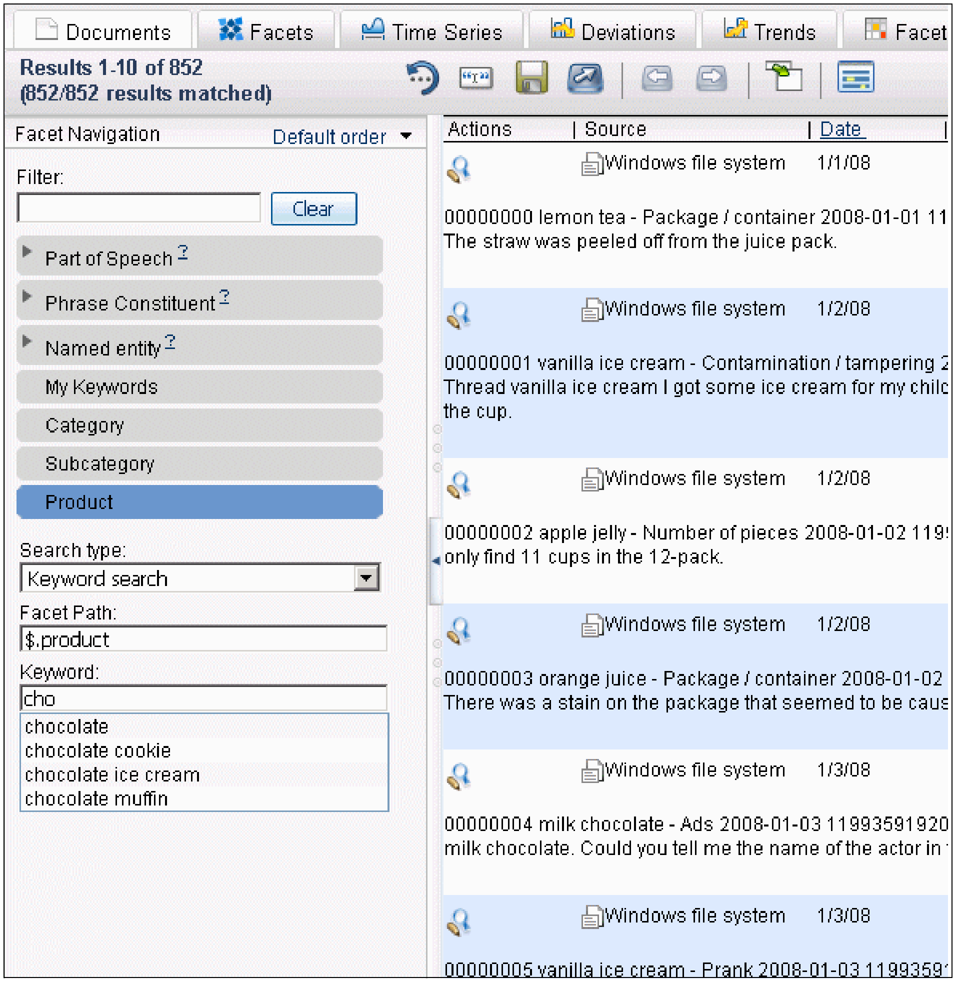
pine juice yesterday	Previous queries	(Estimated results)	3
pine			40
pieces	Index terms	(Estimated results)	100
pine juice			40
piece			40
pieces Shortage chocolate ice cream			10
pieces Shortage chocolate			10
pieces Shortage vanilla ice cream			10
pieces Shortage pastry			10
pieces Shortage milk chocolate			7

Figure 5-10 Query suggestions when you type more than one character

Using the type-ahead feature in the Facet Navigation pane

In addition to using the type-ahead feature in the search query field, you can also use it in the Facet Navigation pane when it is enabled. In the Facet Navigation pane, you can use the type-ahead feature by selecting a facet and typing a keyword.

For example, Figure 5-11 on page 160 shows the result of clicking the **Product** facet, selecting **Keyword** in the Search type field, and typing `cho` in the Keyword field. The possible keywords (up to 10 keywords) for the Product facet that begin with `cho` are displayed in alphabetical order.



Results 1-10 of 852
(852/852 results matched)

Filter:

Search type: Keyword search

Facet Path: $.product

Keyword: cho

chocolate
chocolate cookie
chocolate ice cream
chocolate muffin

Actions	Source	Date
	Windows file system	1/1/08

00000000 lemon tea - Package / container 2008-01-01 11
The straw was peeled off from the juice pack.

Actions	Source	Date
	Windows file system	1/2/08

00000001 vanilla ice cream - Contamination / tampering 2
Thread vanilla ice cream I got some ice cream for my chilc
the cup.

Actions	Source	Date
	Windows file system	1/2/08

00000002 apple jelly - Number of pieces 2008-01-02 119!
only find 11 cups in the 12-pack.

Actions	Source	Date
	Windows file system	1/2/08

00000003 orange juice - Package / container 2008-01-02
There was a stain on the package that seemed to be caus

Actions	Source	Date
	Windows file system	1/3/08

00000004 milk chocolate - Ads 2008-01-03 11993591920
milk chocolate. Could you tell me the name of the actor in

Actions	Source	Date
	Windows file system	1/3/08

00000005 vanilla ice cream - Prank 2008-01-03 1199359

Figure 5-11 Type-ahead feature in the Keyword field of the Facet Navigation pane

To use the type-ahead feature in the Facet Navigation pane, you must select **Keyword search** for the Search type field. Otherwise, if the Search type field is set to Facet search, the Keyword field is not editable.

Type-ahead feature in the Facet Navigation pane: The type-ahead feature in the search box differs from the type-ahead feature in the Facet Navigation pane. The Facet Navigation pane has the following functionality differences:

- Up to 10 suggestions are shown by default, and the number of suggestions to display cannot be changed to another number.
- The suggestions are displayed in alphabetical order, not in frequency order.
- A facet must be selected because the suggestions are based on the facet keyword value.

5.2.6 Saved searches

During your analysis, you can save the current state of your query at anytime by clicking the **Save** icon. When you click the right arrow of the **Saved Searches** tab in the search field area, you see the number of the saved queries and a list of saved queries with the names that you assigned as shown in Figure 5-12.

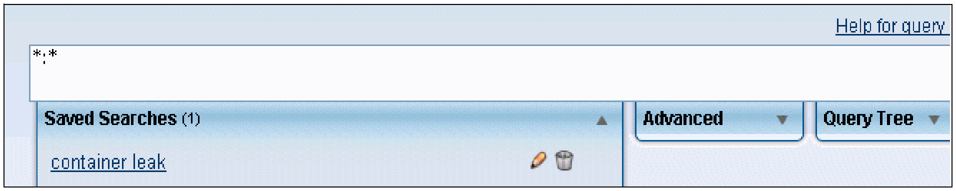
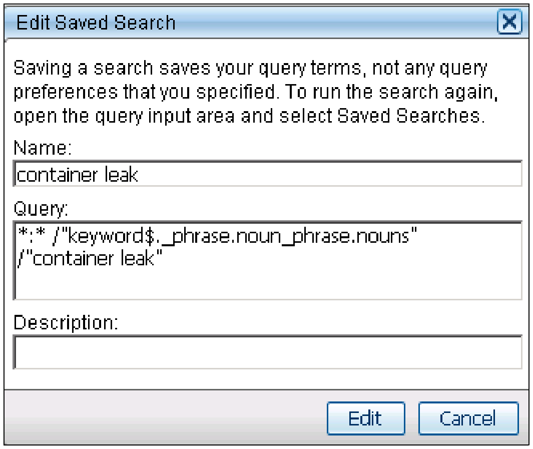


Figure 5-12 Saved Searches window

When you click the name of the saved search, the query starts. If you need to edit the saved query, click the **Pencil** icon on the right side.

In the Edit Saved Search window (Figure 5-13), you can edit the name, query, and description similar to when you saved the query condition before.



Saving a search saves your query terms, not any query preferences that you specified. To run the search again, open the query input area and select Saved Searches.

Name: container leak

Query: *:* /"keyword$._phrase.noun_phrase.nouns" /"container leak"

Description:

Figure 5-13 Edit Saved Search window

You can also click the **Trash** icon to discard the saved query condition.

Global security and saved search: If the login security *is* enabled and you log in to the text miner application, you can save search queries persistently between login sessions. However, you cannot share the query condition with other users.

If the login security is *not* enabled and you are not requested to log in to the text miner application, a saved query is only saved during the current login session. If you need your query condition to be saved longer than your session, you must enable global security as explained in Appendix A, “Security in IBM Content Analytics” on page 633.

5.2.7 Advanced search

On the **Advanced Search** tab, you must first select whether you want to perform a new search or add to a search. The default is to create a search.

Next, you can set the query keywords in each field based on your requirement. For example, you can complete the All of these words field, The exact phrase field, Any of these words field, or None of the words field. You can also specify the start date or end date to search documents within a specified time period.

Figure 5-14 shows the expanded Advanced Search page.

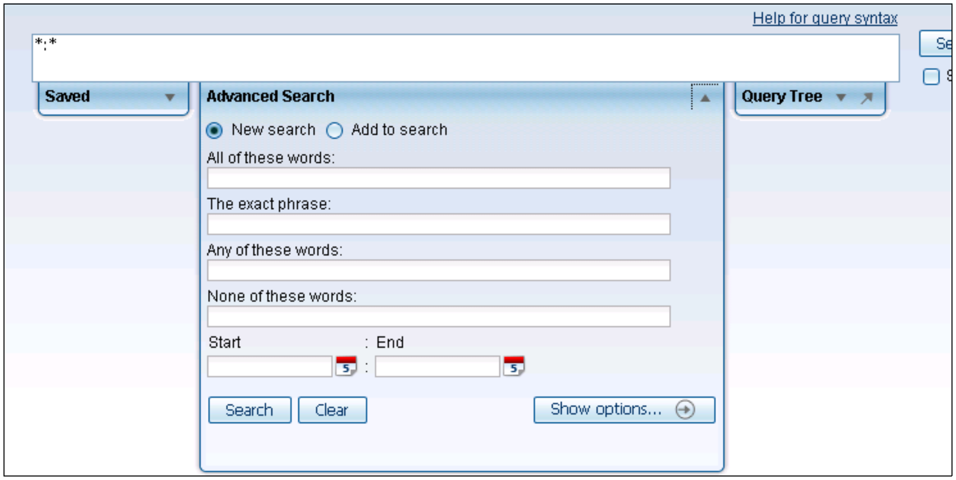


Figure 5-14 Advanced Search window

Show Options button: You can change the Search preference when you click **Show Options** on the Advanced Search page. This option is usually not used in the text miner application.

5.3 Query tree

The query tree is a visual representation of the logical query structure that you enter into the search query field. It helps you to view the logical hierarchical structure of the current query. The query tree is most useful when the entire query becomes large. You can see how the various sections of the query contribute to the overall result set. For each query section, the query tree shows the number of documents that match the specified criteria within the collection.

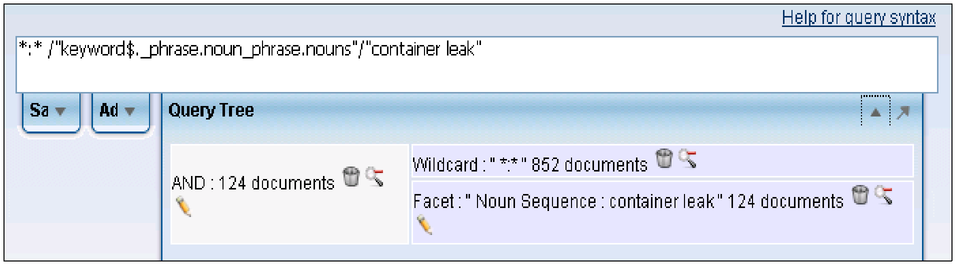
You can change the operator or keywords or tentatively remove each selected node. To remove the selected node, select the **Not And** icon to see the search results without the node included in the query. You can also delete sections of the query by selecting the node that you want to remove and clicking the **Trash** icon next to it.

This section provides information about how to interpret the query tree and use the search operators contained therein.

5.3.1 Accessing the query tree

To view the query tree, select the **Query Tree** tab in the search field area. Each query keyword and search operator is presented as a node.

Figure 5-15 shows the query tree when you narrow your search to “`leak`” using the faceted search with the AND operator to the default query `*:*`.



: /"keyword$._phrase.noun_phrase.nouns"/"container leak"

AND : 124 documents

Wildcard : " *:* " 852 documents

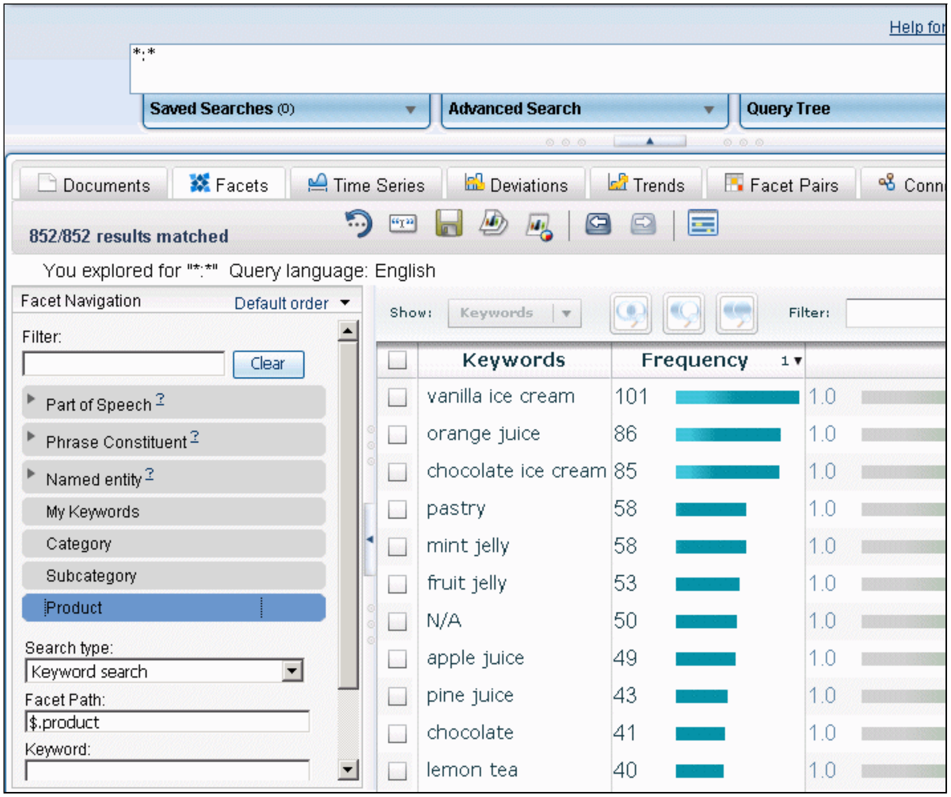
Facet : " Noun Sequence : container leak" 124 documents

Figure 5-15 Query Tree tab

You can minimize the query tree by clicking the triangle icon in the upper-right corner of the Query Tree view. You can maximize the query tree area by clicking the arrow in the upper-right corner of the Query Tree view.

5.3.2 Understanding the query tree

To help you understand the query tree, consider the example shown in Figure 5-16, which shows the results of searching all documents in the collection. By default, the query keyword is set as `*:*`, which means to search for all documents in the collection.



:

852/852 results matched

You explored for "*:*" Query language: English

Search type: Keyword search

Facet Path: $.product

Keyword:

Show: Keywords

Keywords	Frequency	1 ▾
vanilla ice cream	101	1.0
orange juice	86	1.0
chocolate ice cream	85	1.0
pastry	58	1.0
mint jelly	58	1.0
fruit jelly	53	1.0
N/A	50	1.0
apple juice	49	1.0
pine juice	43	1.0
chocolate	41	1.0
lemon tea	40	1.0

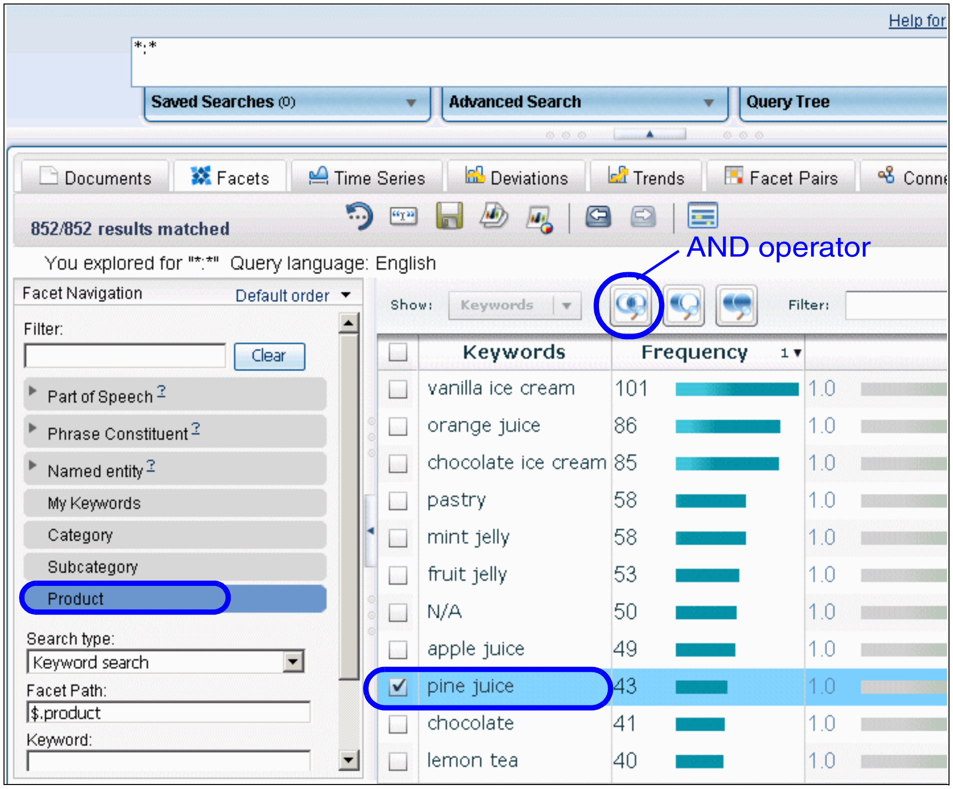
*Figure 5-16 The result with the default query *:* in the Facets view*

5.3.3 Query tree examples

This section shows examples of using the search operators.

The AND operator

From the Facet Navigation pane (Figure 5-17), we select the **Product** facet and the keyword **pine juice**, and then we click the **AND** operator. The results are limited to show only the pine juice-related documents.



:

852/852 results matched

You explored for "*:*" Query language: English

Filter:

Search type: Keyword search

Facet Path: $.product

Keyword:

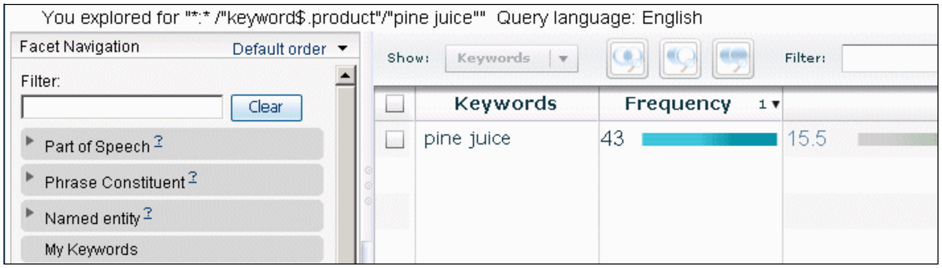
	Keywords	Frequency 1▼	
□	vanilla ice cream	101	1.0
□	orange juice	86	1.0
□	chocolate ice cream	85	1.0
□	pastry	58	1.0
□	mint jelly	58	1.0
□	fruit jelly	53	1.0
□	N/A	50	1.0
□	apple juice	49	1.0
☑	pine juice	43	1.0
□	chocolate	41	1.0
□	lemon tea	40	1.0

Figure 5-17 Selecting the Product facet and pine juice and clicking the AND operator

Figure 5-18 shows the query changes as follows:

```
*:* /"keyword$.product"/"pine juice"
```

Only pine juice-related documents are shown in the Facets view.

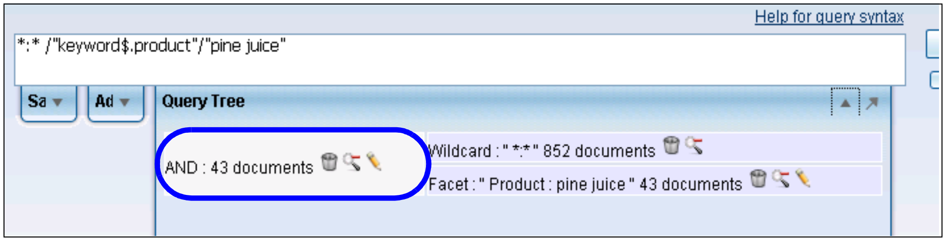


You explored for ""*:* /"keyword$.product"/"pine juice"" Query language: English

□	Keywords	Frequency 1 ▾
□	pine juice	43	15.5

Figure 5-18 The query and result changes in the Facets view with the AND operator

Figure 5-19 shows the associated query tree in this example.



: /"keyword$.product"/"pine juice"

AND : 43 documents

Wildcard : " *:* " 852 documents

Facet : " Product : pine juice " 43 documents

*Figure 5-19 Query tree: The Product facet, pine juice, and *.**

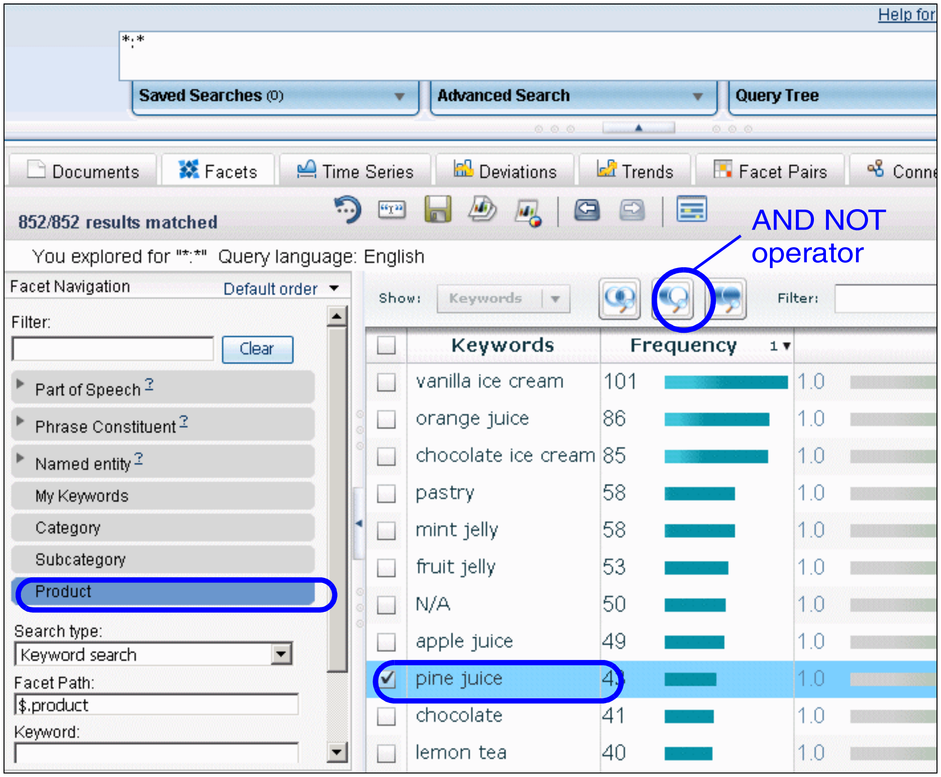
The left node of the query tree shows an AND operator and the number of the documents found (43 documents) as a result of the query following query:

```
*:* /"keyword$.product"/"pine juice"
```

In this case, the query `*.*` and the query `/"keyword$.product"/"pine juice"` are logically aggregated (ANDed). Also Content Analytics finds 43 documents that satisfy the query. From each node on the right side of the query tree, the query `*.*` returns 852 documents, and the facet search returns 43 documents.

The AND NOT operator

From the Facet Navigation pane (Figure 5-20), we select the **Product** facet and **pine juice**, and then we click the **AND NOT** operator.



:

852/852 results matched

You explored for "*:*" Query language: English

Search type: Keyword search

Facet Path: $.product

Keyword:

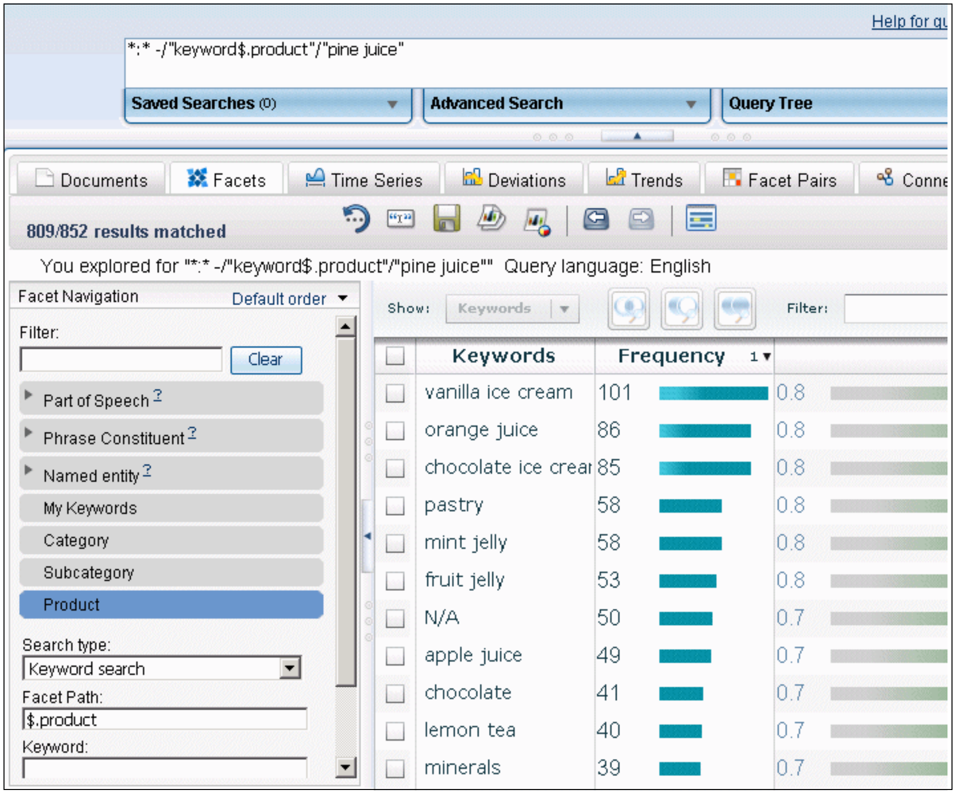
□	Keywords	Frequency 1▼	
□	vanilla ice cream	101	1.0
□	orange juice	86	1.0
□	chocolate ice cream	85	1.0
□	pastry	58	1.0
□	mint jelly	58	1.0
□	fruit jelly	53	1.0
□	N/A	50	1.0
□	apple juice	49	1.0
☑	pine juice	4	1.0
□	chocolate	41	1.0
□	lemon tea	40	1.0

Figure 5-20 Selecting all products, except pine juice, and clicking the AND NOT operator

As a result, the query changes as follows:

```
*:* -/"keyword$.product"/"pine juice"
```

You see the various keywords other than “`pine juice`” in the Facets view, as shown in Figure 5-21. The result between Figure 5-18 on page 167 and Figure 5-21 is different, because this time, we select all products except pine juice.



: -/"keyword$.product"/"pine juice"

809/852 results matched

You explored for "*:* -/"keyword$.product"/"pine juice"" Query language: English

Search type: Keyword search

Facet Path: $.product

Keyword:

□	Keywords	Frequency 1▼	
□	vanilla ice cream	101	0.8
□	orange juice	86	0.8
□	chocolate ice crear	85	0.8
□	pastry	58	0.8
□	mint jelly	58	0.8
□	fruit jelly	53	0.8
□	N/A	50	0.7
□	apple juice	49	0.7
□	chocolate	41	0.7
□	lemon tea	40	0.7
□	minerals	39	0.7

Figure 5-21 All products selected, except pine juice in the Facets view

Figure 5-22 shows the associated query tree in this example.

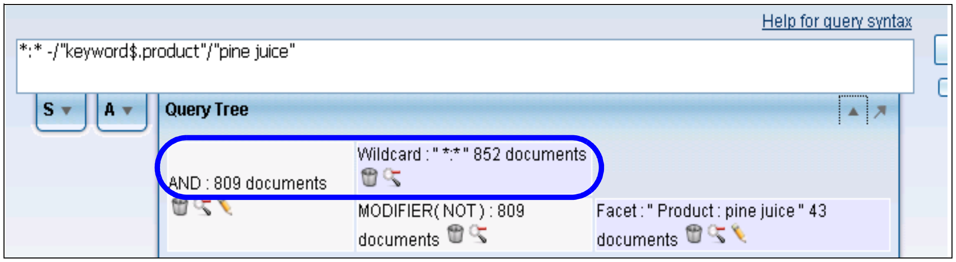

Help for query syntax
: -/"keyword$.product"/"pine juice"
S
A
Query Tree
AND : 809 documents
Wildcard : " *:* " 852 documents
MODIFIER(NOT) : 809 documents
Facet : " Product : pine juice " 43 documents

Figure 5-22 Query Tree: All products selected, except pine juice

In the left node of the query tree, you see an AND operator and the number of the documents found (809 documents) as a result of the following query:

```
*:* -/"keyword$.product"/"pine juice"
```

The NOT operator applied to the search query `/"keyword$.product"/"pine juice"` returns 809 documents, and the query `*.*` returns 852 documents. These results are aggregated with the AND operator, and Content Analytics returns 809 documents.

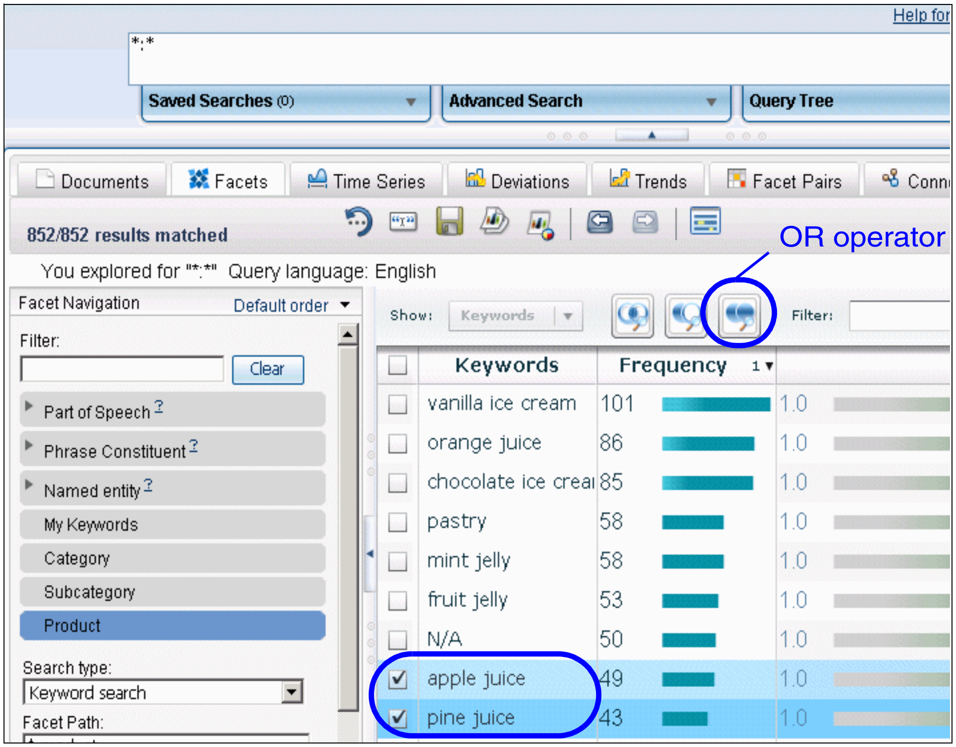
In summary, the following results occurred:

- The default query `*.*` returned 852 documents, which is the total number of documents in the collection.
- The query that limits the results to pine juice only (with `AND pine juice`) returned 43 documents.
- The query that excluded pine juice (with `AND NOT pine juice`) returned 809 documents.

The last two results reflect the total number of documents found in the collection.

The OR operator

The example in this section shows how many documents are returned, including the apple juice or pine juice product-related information. From the Facet Navigation window (Figure 5-23), we select the **Product** facet and both **apple juice** and **pine juice**. Then we click the **OR** operator.



:

852/852 results matched

You explored for "*:*" Query language: English

Filter:

Search type: Keyword search

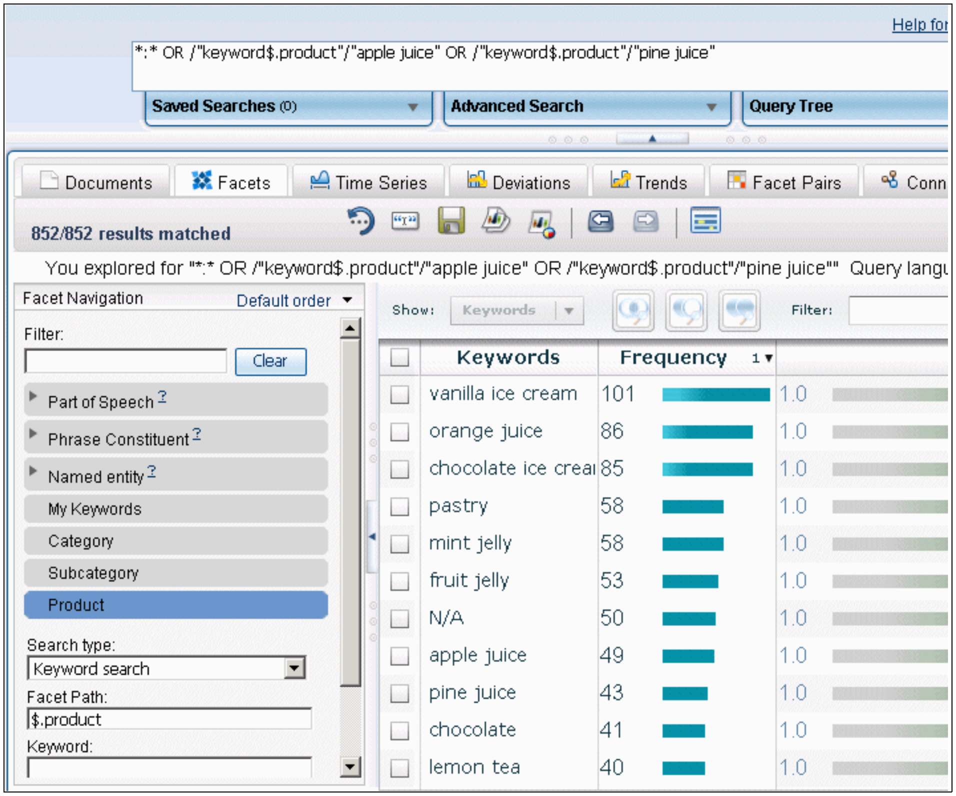
□	Keywords	Frequency 1 ▾	
□	vanilla ice cream	101	1.0
□	orange juice	86	1.0
□	chocolate ice creai	85	1.0
□	pastry	58	1.0
□	mint jelly	58	1.0
□	fruit jelly	53	1.0
□	N/A	50	1.0
☑	apple juice	49	1.0
☑	pine juice	43	1.0

Figure 5-23 Selecting apple juice or pine juice and clicking the OR operator

This operation changes the query as follows:

```
*:* OR /"keyword$.product"/"apple juice" OR /"keyword$.product"/
"pine juice"
```

The query returns 852 documents, which is the same result if you search with the query `*:*` as shown in the Figure 5-24.



: OR /"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"

852/852 results matched

You explored for "*:* OR /"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"" Query langu

Filter:

- Part of Speech ?
- Phrase Constituent ?
- Named entity ?
- My Keywords
- Category
- Subcategory
- Product

Search type: Keyword search

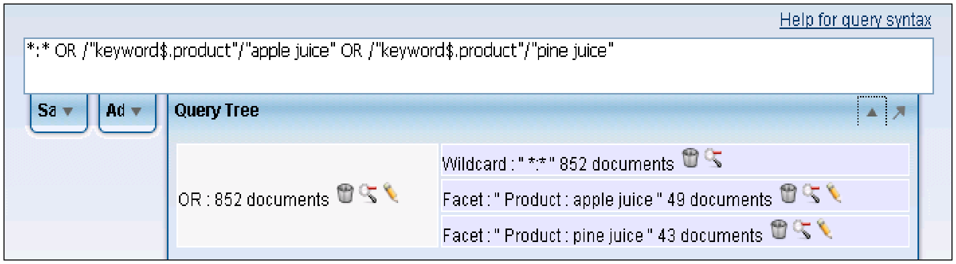
Facet Path: $.product

Keyword:

□	Keywords	Frequency 1▼	
□	vanilla ice cream	101	1.0
□	orange juice	86	1.0
□	chocolate ice crea	85	1.0
□	pastry	58	1.0
□	mint jelly	58	1.0
□	fruit jelly	53	1.0
□	N/A	50	1.0
□	apple juice	49	1.0
□	pine juice	43	1.0
□	chocolate	41	1.0
□	lemon tea	40	1.0

Figure 5-24 Apple juice or pine juice, using the OR operator, but all documents returned

Figure 5-25 shows the associated query tree in this example.



: OR /"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"

OR : 852 documents

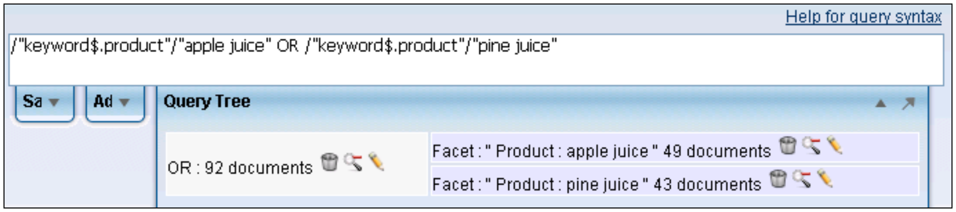
- Wildcard : " *:* " 852 documents
- Facet : " Product : apple juice " 49 documents
- Facet : " Product : pine juice " 43 documents

Figure 5-25 Query tree with the OR operator

As you can see, the OR operator is applied for the selected facet search keywords `apple juice` and `pine juice`. It is also applied to the default query `*:*`.

As long as the default query is included with the OR operator, Content Analytics returns the same result as the default query.

To get the documents that are only apple juice or pine juice related, we must remove the default query `*:*` from the entire query. The easiest way is to remove the node that contains the default query `*:*` from the query tree. After we click the **Trash** icon on the right side of the default query `*:*` node (in Figure 5-25 on page 172), the node is removed from the query. Figure 5-26 shows the new query tree.



/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"

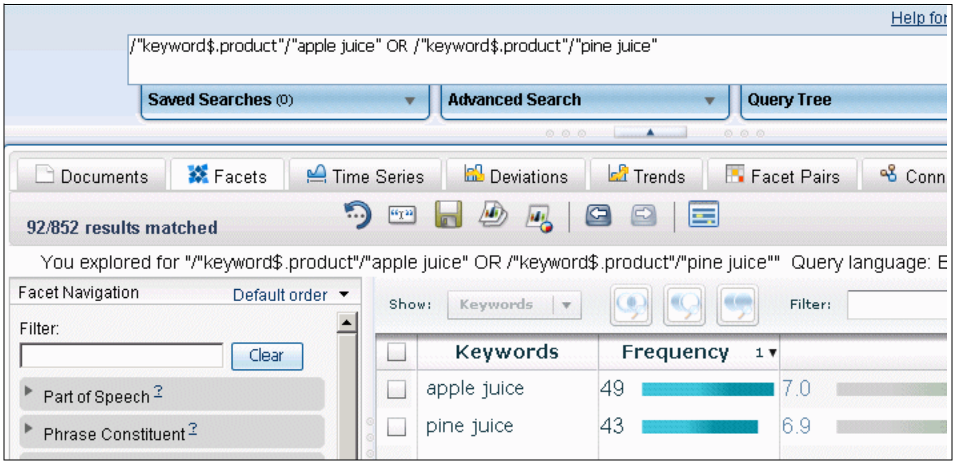
OR : 92 documents

Facet : " Product : apple juice " 49 documents

Facet : " Product : pine juice " 43 documents

Figure 5-26 Query Tree changes after one condition is removed

Now the query changes to apple juice or pine juice only. You no longer see the default query node in the query tree. When we go back to the Facets view, a different result is displayed (Figure 5-27), as compared to the result shown in Figure 5-24 on page 172.



/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"

92/852 results matched

You explored for "/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"" Query language: E

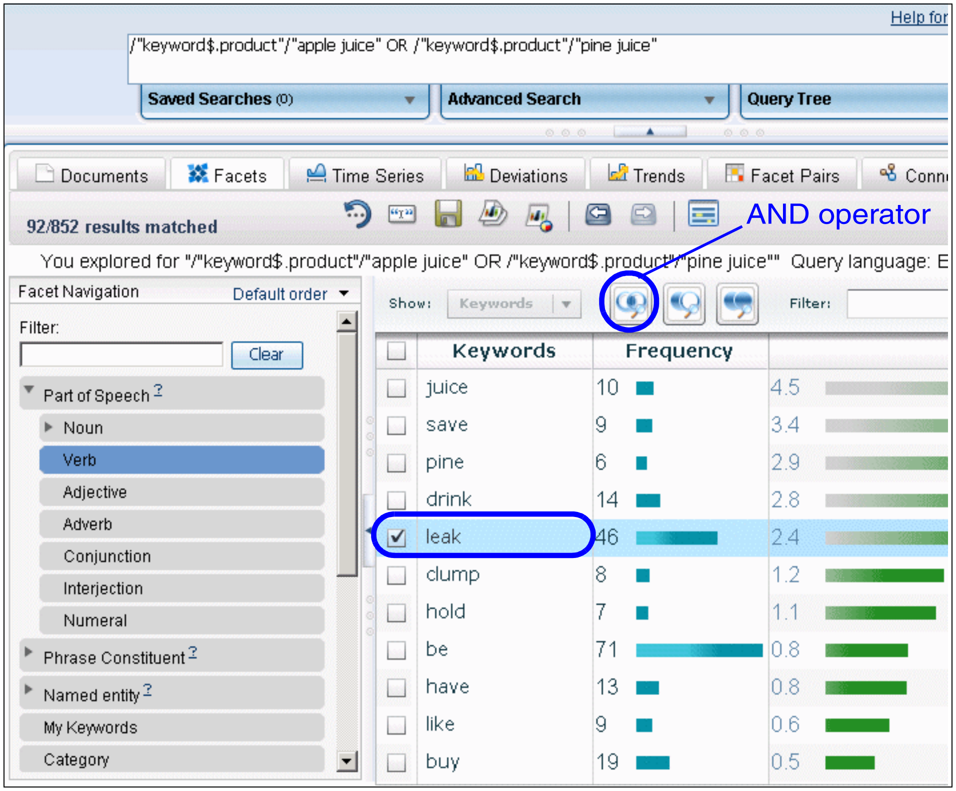
Show: Keywords

□	Keywords	Frequency 1 ▾
□	apple juice	49 7.0
□	pine juice	43 6.9

Figure 5-27 Apple juice or pine juice products using only the OR operator

Applying search operators several times

In this example, we apply the search operators several times and see how the query tree changes. We want to drill down in the search results that we saw in "The OR operator" on page 171 from a different aspect. From the Facet Navigation window (Figure 5-28), we select the **Verb** facet and the keyword **leak**. Then we click the **AND** operator.



/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"

92/852 results matched

You explored for "/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"" Query language: E

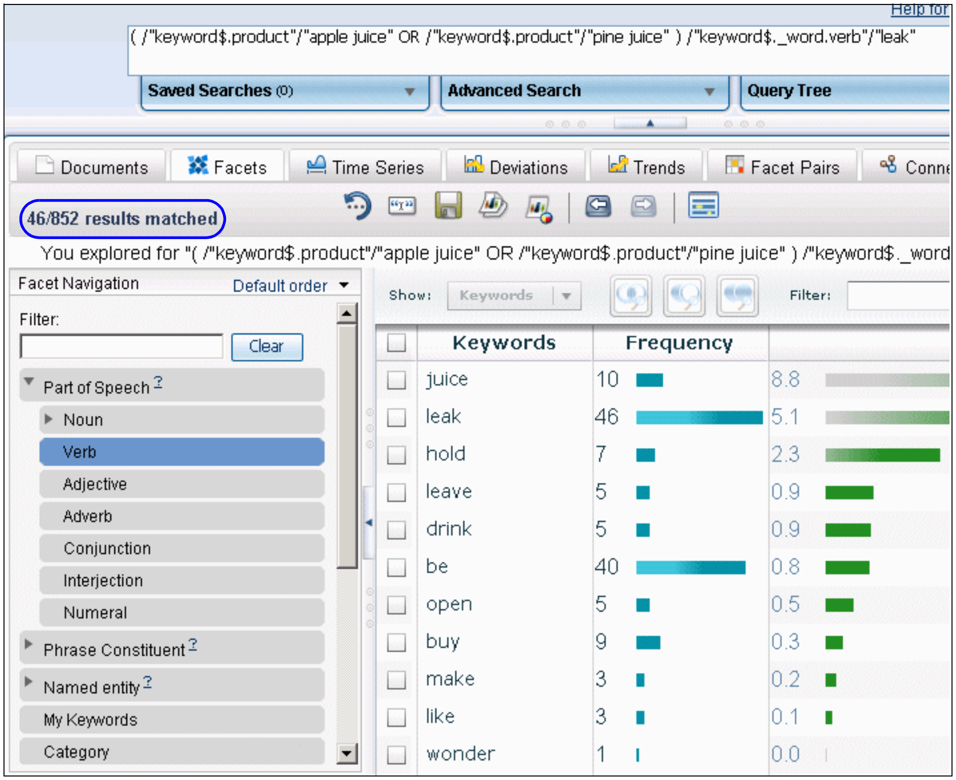
☐	Keywords	Frequency	
☐	juice	10	4.5
☐	save	9	3.4
☐	pine	6	2.9
☐	drink	14	2.8
☑	leak	46	2.4
☐	clump	8	1.2
☐	hold	7	1.1
☐	be	71	0.8
☐	have	13	0.8
☐	like	9	0.6
☐	buy	19	0.5

Figure 5-28 Selecting a keyword and clicking the AND operator button

This operation changes the query to find all documents that have either "apple juice" or "pine juice," and that have the word "leak":

```
(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice")
/"keyword$._word.verb"/"leak"
```

As shown in Figure 5-29, the query returns 46 documents.



(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") /"keyword$._word.verb"/"leak"

46/852 results matched

You explored for "(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") /"keyword$._word

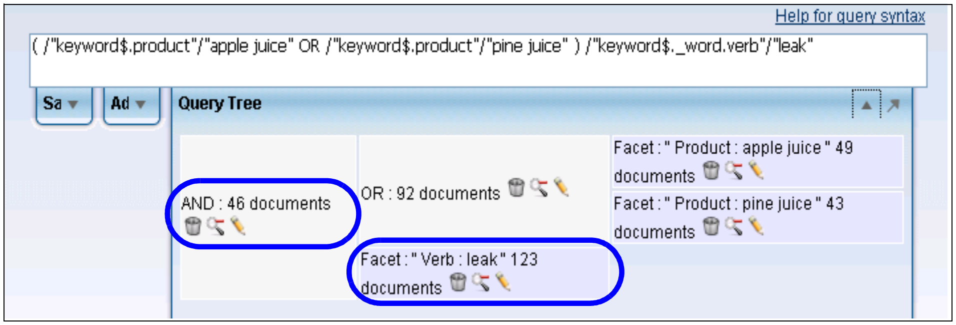
Keywords	Frequency	
juice	10	8.8
leak	46	5.1
hold	7	2.3
leave	5	0.9
drink	5	0.9
be	40	0.8
open	5	0.5
buy	9	0.3
make	3	0.2
like	3	0.1
wonder	1	0.0

Figure 5-29 Combination operators: (Apple juice OR pine juice) AND leak

Notice that the second AND operator is applied at the end of the existing apple juice or pine juice query:

```
/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice"
```

The existing query is set as one group with parentheses. Figure 5-30 shows the associated query tree in this example.



(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") /"keyword$._word.verb"/"leak"

Help for query syntax
Sa
Ad
Query Tree
AND : 46 documents
OR : 92 documents
Facet : " Product : apple juice " 49 documents
Facet : " Product : pine juice " 43 documents
Facet : " Verb : leak " 123 documents

Figure 5-30 Query tree: (Apple juice OR pine juice) AND leak

The AND operator is applied to the node, which returns the selected facet keywords “apple juice” or “pine juice.”

From the query tree, in summary, the following results occur:

- 49 documents for the apple juice product
- 43 documents for the pine juice product
- 92 documents for either apple juice or pine juice products
- 123 documents with a leak problem
- 46 documents with a leak problem that applies to apple juice or pine juice

You can apply the search operators for the selected nodes. In this example, we do not explicitly select the specific node in the query tree. If you do not select the specific node in the query tree, the search operator is applied to the “root” node on the left side. However, if you select a specific node and apply the search operator, the search operator is applied to the selected node. Thus, you can build the complex query further by iterating the process, such as selecting a specific node and applying the search operator several times.

5.3.4 Editing the query tree

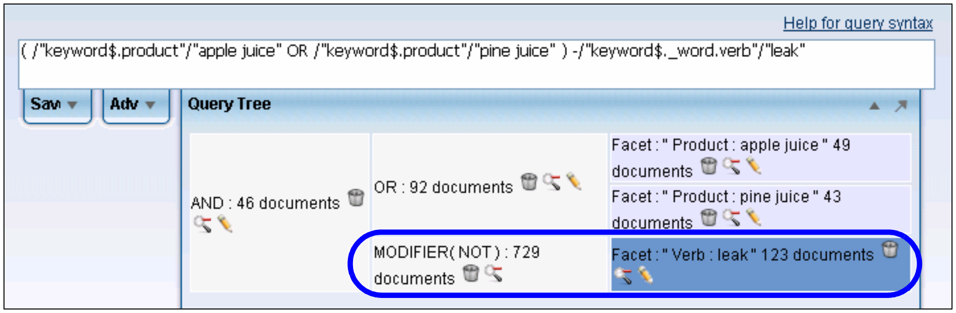
With the query tree, you can also edit the query to find the useful query for your analysis. Table 5-2 defines the icons that are displayed in the query tree.

Table 5-2 The query tree icons and description

icon	Description
Exclude 🔍	You click this icon to exclude a selected term from the query. When you click this icon, the AND NOT operator is added to the query tree. This icon is useful when you quickly examine a query that does not have the term.
Edit ✏	You click this icon to edit the keyword or the BOOLEAN operator. This icon is useful when you change the operator type from AND to OR, and vice versa. You can also edit the keyword in the node.
Delete 🗑	You click this icon to delete the keyword or node from the query tree.

Excluding a node

When you build the query tree, each term is represented as a node. When you want to exclude one of the query terms, click the **Exclude** icon. As a result, the selected node is added with the AND NOT operator. For example, if you click the **Exclude** icon associated with the `/"keyword$._word.verb"/"leak"` node as shown in Figure 5-30 on page 176, the query tree changes as shown in Figure 5-31.


(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") -/"keyword$._word.verb"/"leak"

AND : 46 documents

OR : 92 documents

Facet : " Product : apple juice " 49 documents

Facet : " Product : pine juice " 43 documents

MODIFIER(NOT) : 729 documents

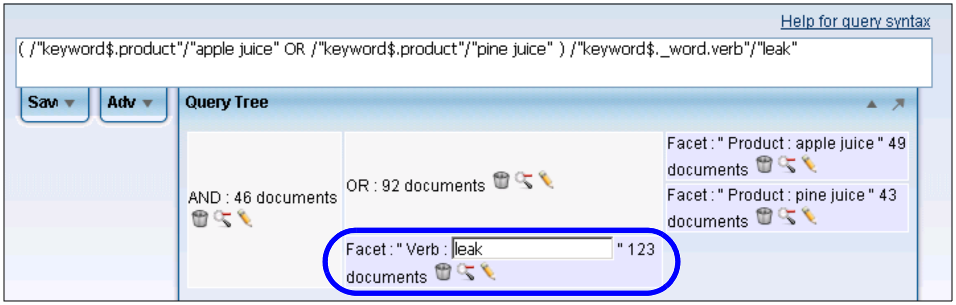
Facet : " Verb : leak " 123 documents

Figure 5-31 Excluding a keyword node in the Query Tree

The search results are immediately displayed in your view.

Editing a node

You can edit the operator type or the keyword without building the query from the beginning. When you click the **Edit** icon for a keyword node, a selection list is displayed, as shown in Figure 5-32.



(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") /"keyword$._word.verb"/"leak"

AND : 46 documents

OR : 92 documents

Facet : " Product : apple juice " 49 documents

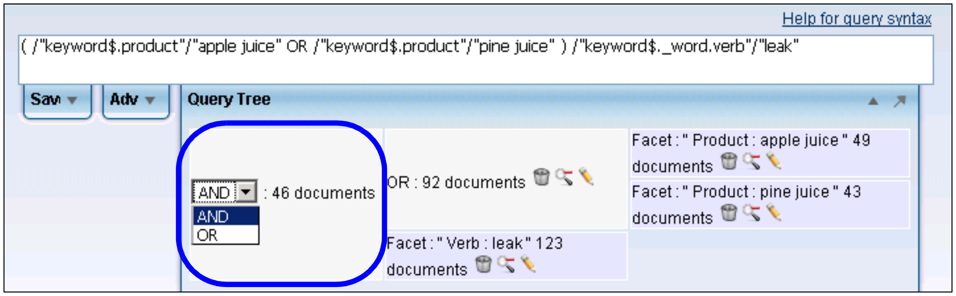
Facet : " Product : pine juice " 43 documents

Facet : " Verb : leak " 123 documents

Figure 5-32 Editing a keyword node in the Query Tree

For example, you can modify the keyword "`leak`" to be a different keyword found in the Verb facet.

You can change the search operator by using a drop-down field, as shown in Figure 5-33.



(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") /"keyword$._word.verb"/"leak"

AND : 46 documents

OR : 92 documents

Facet : " Product : apple juice " 49 documents

Facet : " Product : pine juice " 43 documents

Facet : " Verb : leak " 123 documents

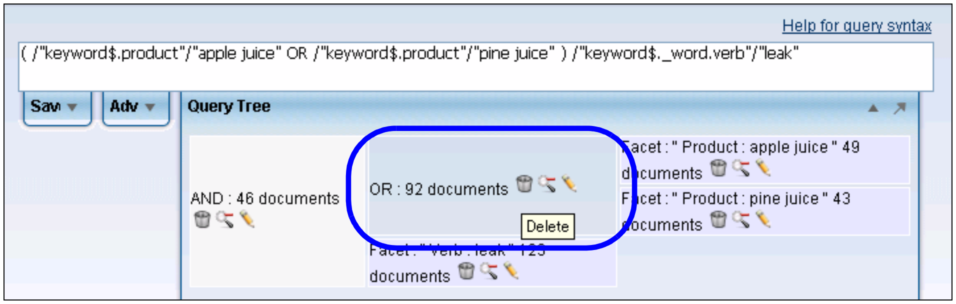
Figure 5-33 Editing search operator node in the Query Tree

The query change is reflected immediately and the search results are updated in your view. You can review the query results for various queries without building the query from the scratch.

Deleting a node

When you want to delete a keyword, click the **Delete** icon next to the particular keyword. This operation deletes the selected node and its child nodes (if they exist). Make sure that you do not use the keyword anymore before you delete it.

For example, consider when you delete the OR node as shown in Figure 5-34.

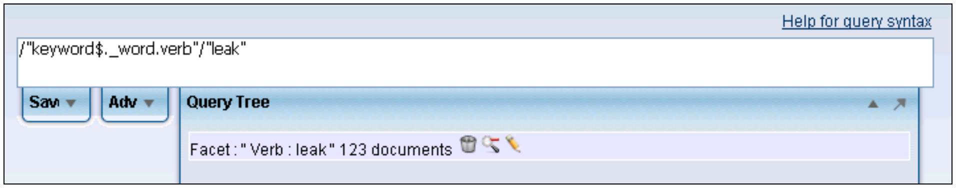


(/"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice") /"keyword$._word.verb"/"leak"

Help for query syntax
Sav
Adv
Query Tree
AND : 46 documents
OR : 92 documents
Delete
Facet : " Product : apple juice " 49 documents
Facet : " Product : pine juice " 43 documents
Facet : " Verb : leak " 123 documents

Figure 5-34 Deleting the search operator node in the Query Tree

In this case, the child nodes (`/"keyword$.product"/"apple juice"` and `/"keyword$.product"/"pine juice"`) of the selected OR node are deleted, as shown in Figure 5-35.



/"keyword$._word.verb"/"leak"

Figure 5-35 After the search operator node is deleted in the Query Tree

If you want to exclude a keyword from the query, click the **Exclude** icon next to that particular keyword. Excluding a node in the query tree can be used instead of deleting the node.

The query tree is helpful when you analyze the data with queries, especially when you examine the data from different aspects. You can modify the query by clicking icons on the right side to correct the query. The change is reflected immediately to your query statement, and you can continue editing the query until you find the useful query for your analysis.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *searching collections and using the query tree*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

5.4 Query builder

You can see and build your query from the query tree, but the query builder helps you to build complex queries. The query builder helps you to build the query by selecting keywords or facets for the selected documents and adding the built query to an existing query or selected query node in the query tree.

In addition, a preview of the selected documents is displayed with the query builder. You can quickly select keywords from the preview, add the condition to the existing query, and verify the results immediately. After the query is built with the query tree, you can interact with the query builder to confirm the structure of the query, and you can modify the query to find further insight.

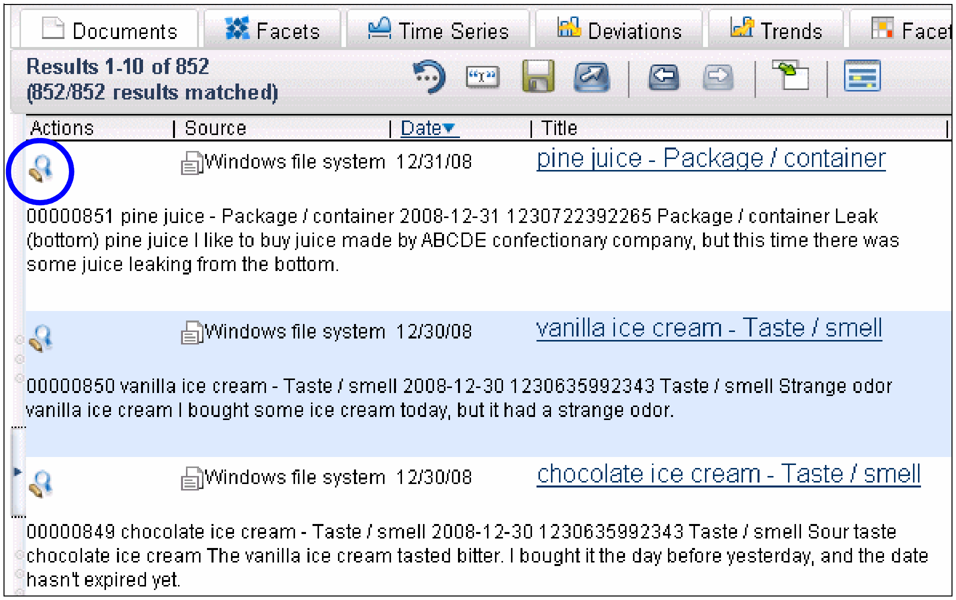
This section explains how to use the query builder and how to interact with the query tree.

5.4.1 Accessing the query builder

To use the Query Builder, the “Build queries with the query builder” application user role must be enabled. By default, this user privilege is not enabled. You must enable the feature from the administration console explicitly.

Configuring application user role: See Appendix A, “Security in IBM Content Analytics” on page 633, for further detail.

After you enable the query builder feature, the Query Builder icon is displayed in the Actions column within the Documents view, as shown in Figure 5-36.



Results 1-10 of 852
(852/852 results matched)

Actions	Source	Date▼	Title
	Windows file system	12/31/08	pine juice - Package / container
	Windows file system	12/30/08	vanilla ice cream - Taste / smell
	Windows file system	12/30/08	chocolate ice cream - Taste / smell

00000851 pine juice - Package / container 2008-12-31 1230722392265 Package / container Leak (bottom) pine juice I like to buy juice made by ABCDE confectionary company, but this time there was some juice leaking from the bottom.

00000850 vanilla ice cream - Taste / smell 2008-12-30 1230635992343 Taste / smell Strange odor vanilla ice cream I bought some ice cream today, but it had a strange odor.

00000849 chocolate ice cream - Taste / smell 2008-12-30 1230635992343 Taste / smell Sour taste chocolate ice cream The vanilla ice cream tasted bitter. I bought it the day before yesterday, and the date hasn't expired yet.

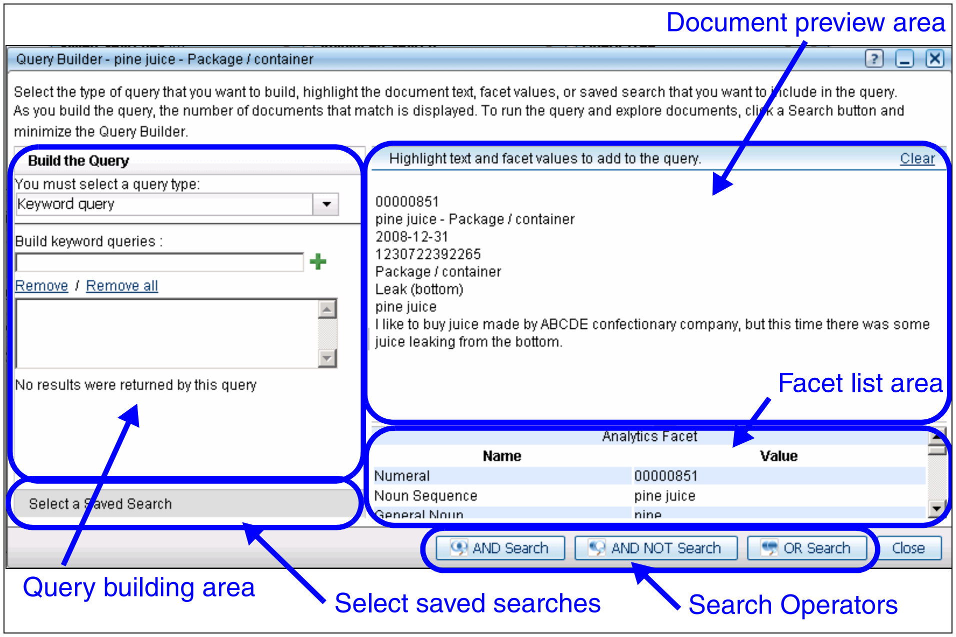
Figure 5-36 Query Builder icon appears in Actions column

When you click the **Actions** icon, a Query Builder window (Figure 5-37 on page 182) opens.

5.4.2 Features of the Query Builder window

The Query Builder window has the following main areas (Figure 5-37):

- Query building area
- Document preview area
- Facet list area



Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

You must select a query type:

No results were returned by this query

00000851
pine juice - Package / container
2008-12-31
1230722392265
Package / container
Leak (bottom)
pine juice
I like to buy juice made by ABCDE confectionary company, but this time there was some juice leaking from the bottom.

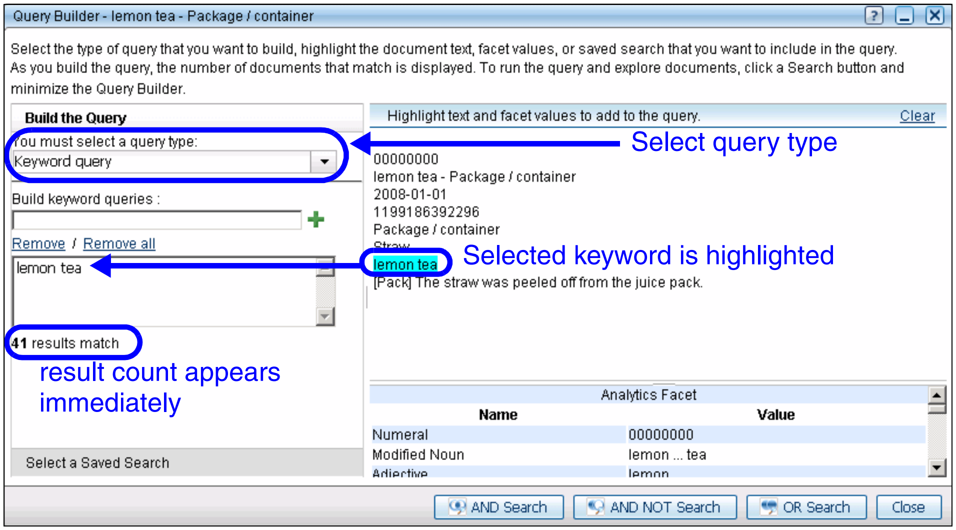
Name	Value
Numeral	00000851
Noun Sequence	pine juice
General Noun	pine

Figure 5-37 Query Builder main window

You can build the following query types from the query builder:

- Keyword query
- Phrase query
- Facet query
- Field query
- Date query
- Parametric range query
- Proximity query
- Fuzzy query
- Boost query
- Exact match query
- Base form match query

You select the query type and drag the keyword or facet in the document from the right area of the query builder (either from the document preview area or the facet list area). Then, as shown in Figure 5-38, the query keyword or facet is added immediately, and you see the result count.



Query Builder - lemon tea - Package / container

Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

Build the Query

You must select a query type:

Keyword query

Build keyword queries :

Remove / Remove all

lemon tea

41 results match

Select a Saved Search

Highlight text and facet values to add to the query.

Clear

00000000
lemon tea - Package / container
2008-01-01
1199186392296
Package / container
Straw
lemon tea
[Pack] The straw was peeled off from the juice pack.

Analytics Facet

Name	Value
Numeral	00000000
Modified Noun	lemon ... tea
Adjective	lemon

Select query type

Selected keyword is highlighted

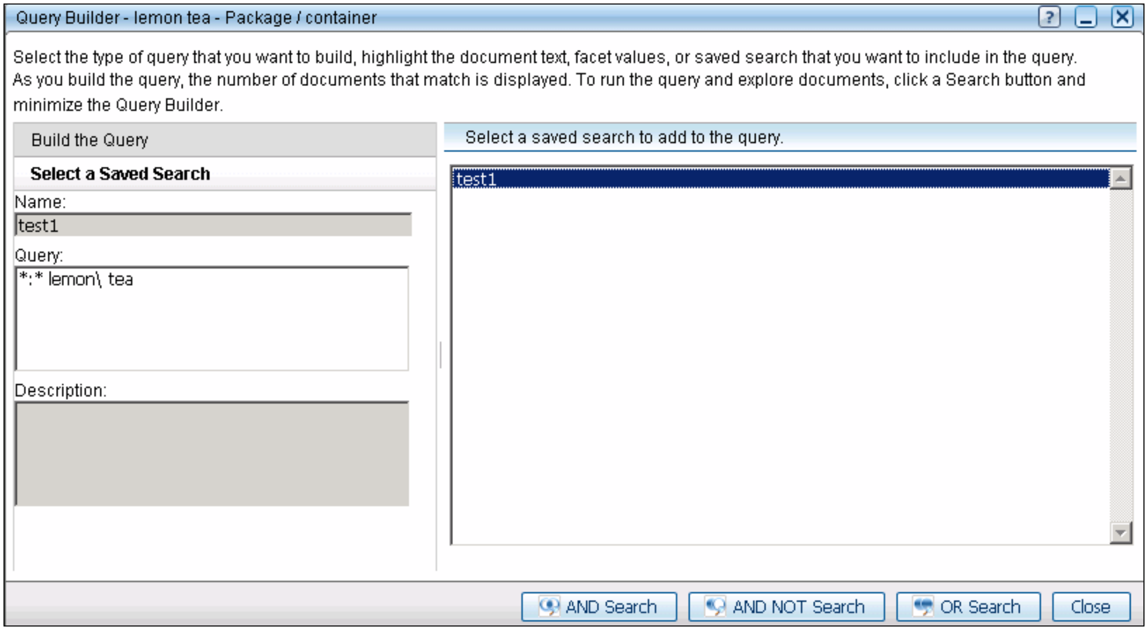
result count appears immediately

Figure 5-38 Selecting keywords in the document on the query builder

Result count in the query building area: The result count in the query builder is the result of the current query that you built using the query builder. The query that you already issued in the text miner application (such as in the Documents view) is not considered at this point.

You can use saved searches to build the query. In this case, you click the **Select a Saved Search** area below the query building area, as shown in Figure 5-37 on page 182.

As shown in Figure 5-39 on page 184, the query builder shows the list of saved searches. The saved query is displayed in the Query area. You can use the saved query or add the saved query to the selected node in the query tree with the search operator.



Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

Name: test1

Query: *:* lemon\ tea

Description:

Select a saved search to add to the query.

test1

Figure 5-39 Selecting a saved searches in the query builder

When you decide on the query to be examined, select the search operators (AND, AND NOT, or OR) at the bottom of the Query Builder window. Then your search results are updated. After you select the Boolean operator, the Query Builder window is minimized as shown in Figure 5-40. You can open the Query Builder window again by selecting the **Expand this area** icon.

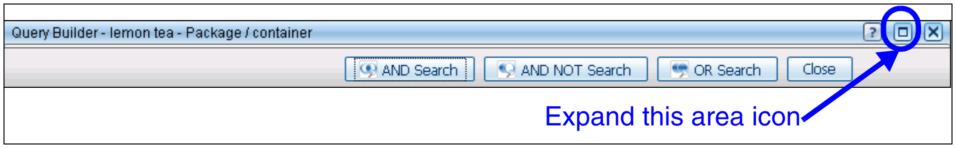
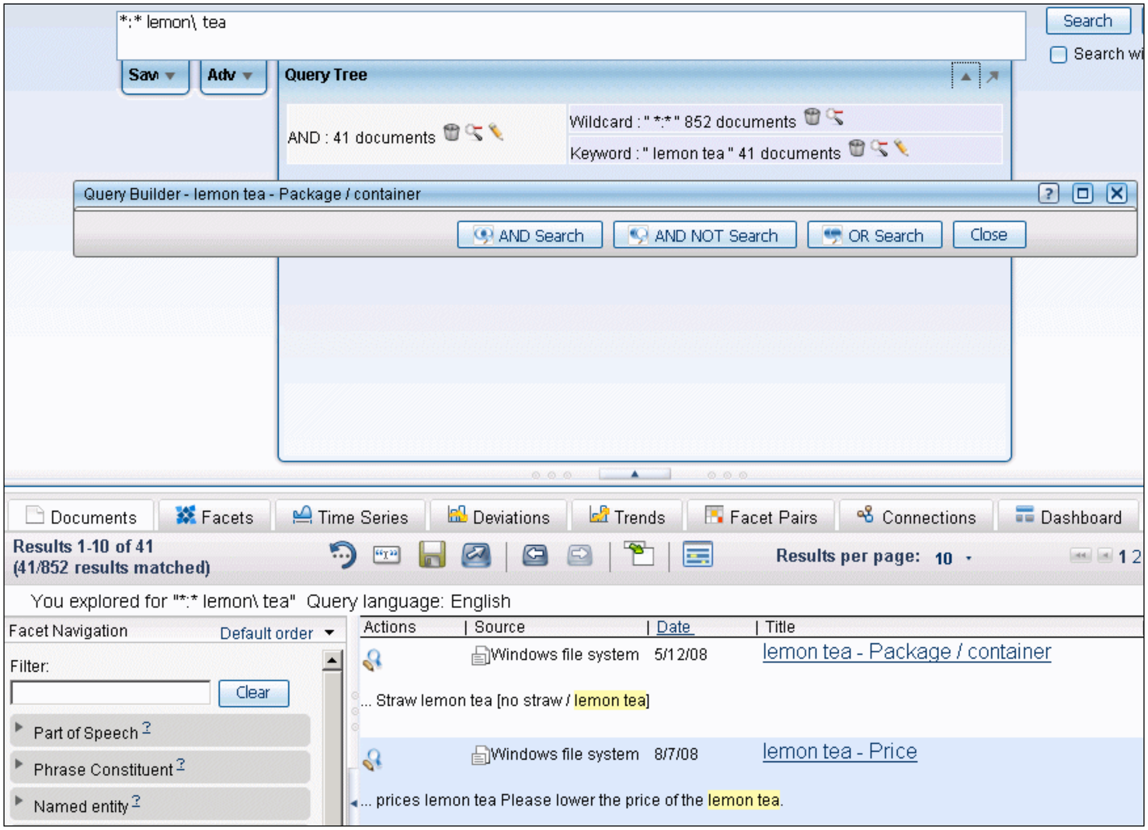


Figure 5-40 After the query is issued from the query builder

In the search box area, the entire query and the query tree structure on the Query Tree tab are displayed, as shown in Figure 5-41. The selected keyword is highlighted in the summary of each document in the Documents view.



: lemon\ tea

AND : 41 documents

Wildcard : " *:* " 852 documents

Keyword : " lemon tea " 41 documents

Query Builder - lemon tea - Package / container

Results 1-10 of 41
(41/852 results matched)

You explored for "*:* lemon\ tea" Query language: English

Actions	Source	Date	Title
	Windows file system	5/12/08	lemon tea - Package / container
... Straw lemon tea [no straw / lemon tea]			
	Windows file system	8/7/08	lemon tea - Price
... prices lemon tea Please lower the price of the lemon tea.			

Figure 5-41 The query with the query tree and selected keyword highlighted in the Documents view

After you issue the query, you can go back and forth to the query builder to build more complex queries. You can add the query to the selected node in the query tree, and you can operate the query from the query tree.

5.4.3 Using the query builder

This section shows examples of how to use the query builder. You can build a query with the query builder by using the following steps. Then you repeat the steps until you decide whether the query provides the necessary information for your analysis.

1. Select a document from the Documents view and open the Query Builder window.
2. Build a query in the query builder. Select the query type from the selection box, and select the keyword you are interested in from the Document preview area.
3. Add the query to the current query with one of the search operators and decide whether it is what you want:
   - If you need to modify the query, you can modify the query either from the query builder or from the query tree.
   - When you want to add the built query to a certain node in the query tree, select the query node in the query tree and issue the query with the search operator.
4. Repeat steps 1–3 to build the query that you want. You can use the query tree to help you build the query.
5. After you complete the query building, save the query for reuse in the future, such as for document export or to build another query.

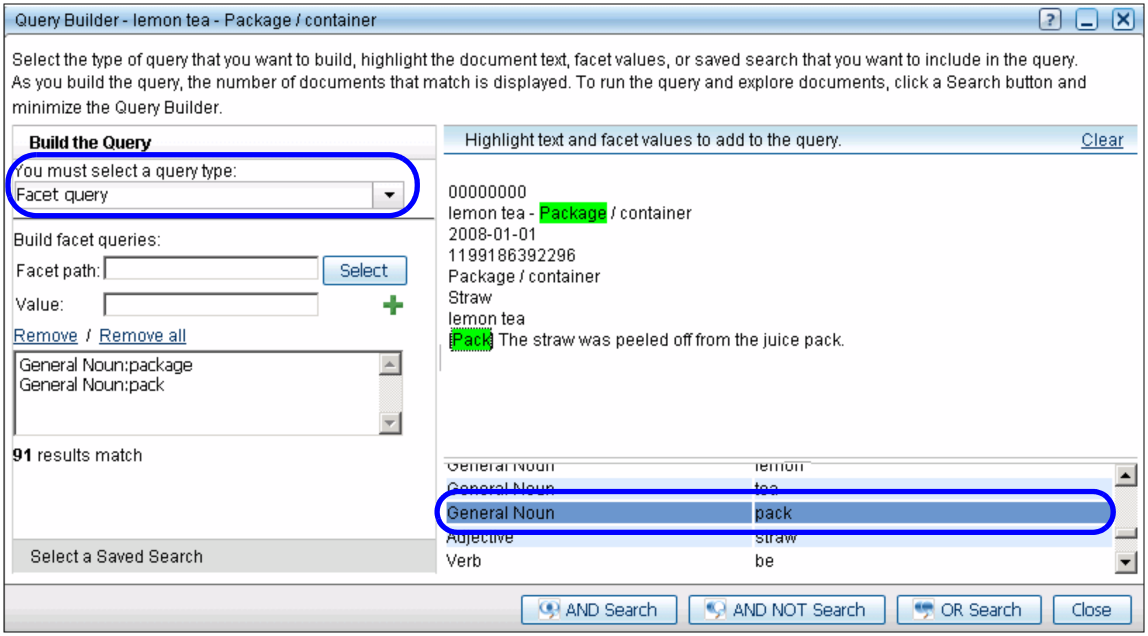
Selecting a document and opening the query builder

Consider an example where a specific document contains keywords that you want to use to find similar information. The document contains information about the “lemon tea” product. Select the document in the Documents view, and open the Query Builder window, as shown in Figure 5-38 on page 183.

Building and issuing a query

After you open the query builder, a preview of the selected document is displayed in the Document preview area or the Facet list area. Select a query type from the drop-down list, and drag the keyword in the document.

In this example, we select **Facet query** as the query type, and select **General Noun: package** from the Facet list. We select another **General Noun: pack** from the Facet list, as shown in Figure 5-42 on page 187.



Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

You must select a query type: Facet query

Build facet queries:

Facet path:

Value:

General Noun:package
General Noun:pack

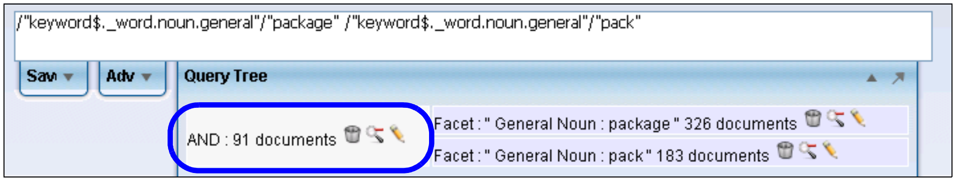
91 results match

00000000
lemon tea - Package / container
2008-01-01
1199186392296
Package / container
Straw
lemon tea
Pack The straw was peeled off from the juice pack.

General Noun	lemon
General Noun	tea
General Noun	pack
Adjective	straw
Verb	be

Figure 5-42   Selecting the Facet query with two keywords from the Facet list

After selecting the keywords, issue the query with the AND search operator. The query results are updated, and 91 documents match this node in the query, as shown in Figure 5-43.



```
/"keyword$._word.noun.general"/"package" /"keyword$._word.noun.general"/"pack"
```

AND : 91 documents
Facet : " General Noun : package " 326 documents
Facet : " General Noun : pack " 183 documents

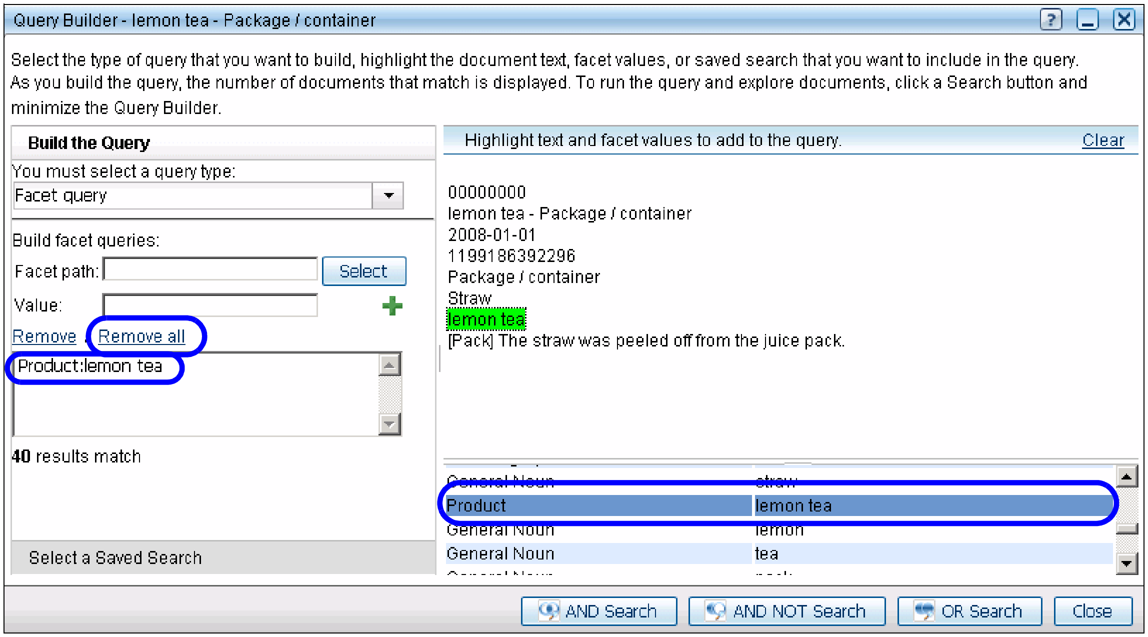
Figure 5-43   The query tree for the query with AND operator

Adding a query to a selected node in the query tree

You can also add a query to the selected node in the query tree to narrow down the data further. When you do not explicitly select the node in the query, the query is added to the entire query.

For example, perform the following steps to add a query to the entire node:

1. Remove the previous condition by clicking the **Remove all** link in the Build the Query area.
2. In the query text area, type `Product: lemon tea` as shown in Figure 5-44.
3. Click **AND NOT Search** to add the Product: lemon tea keyword with the AND NOT search operator. This option shows how many documents are not regarding the lemon tea product.



Query Builder - lemon tea - Package / container

Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

Build the Query

You must select a query type: Facet query

Build facet queries:

Facet path:

Value:

Product:lemon tea

40 results match

Select a Saved Search

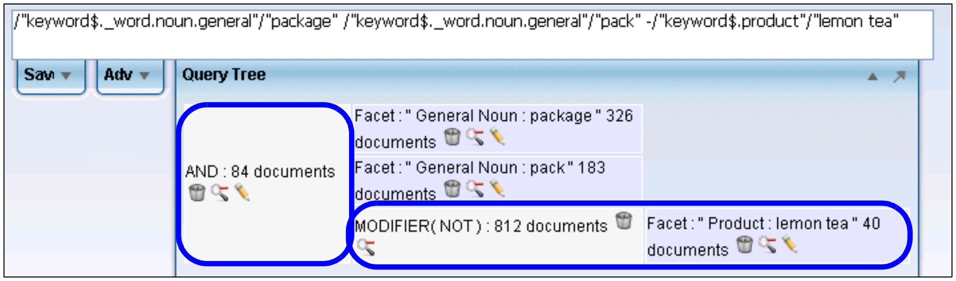
Highlight text and facet values to add to the query.

00000000
lemon tea - Package / container
2008-01-01
1199186392296
Package / container
Straw
lemon tea
[Pack] The straw was peeled off from the juice pack.

General Noun	straw
Product	lemon tea
General Noun	lemon
General Noun	tea

Figure 5-44 Selecting the facet query after flushing the existing keyword

In this case, we do not select any node in the query tree. The query is added to the entire query, and the new query returns 84 documents, as shown in Figure 5-45.



/"keyword$._word.noun.general"/"package" /"keyword$._word.noun.general"/"pack" -/"keyword$.product"/"lemon tea"

AND : 84 documents

Facet : " General Noun : package " 326 documents

Facet : " General Noun : pack " 183 documents

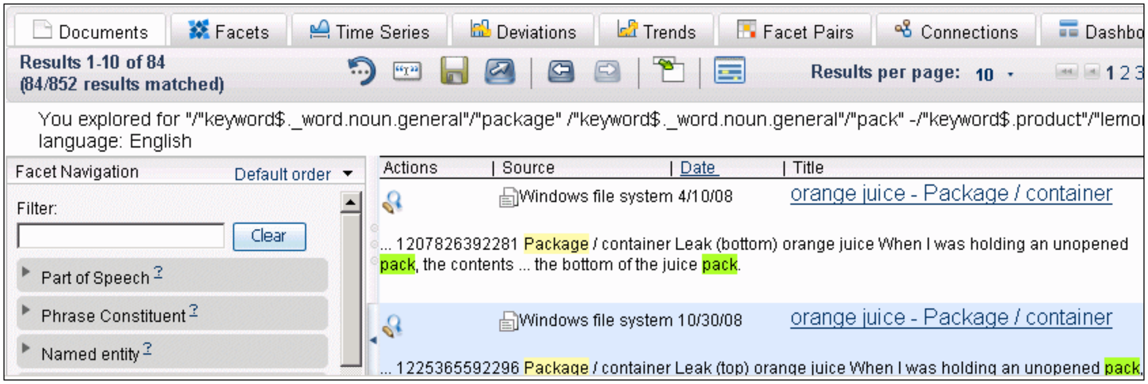
MODIFIER(NOT) : 812 documents

Facet : " Product : lemon tea " 40 documents

Figure 5-45 The query tree for the query with the AND NOT operator

The results indicate that the documents that mention the “package” and “pack” nouns might be related to products other than the lemon tea product. The new query produced a result set of 84 documents, which means that only 7 of 91 documents are related to the lemon tea product.

When we look at the search result in the Documents view, a document that contains the term “orange juice” is displayed at the top of the search result, as shown in Figure 5-46. We wonder if the documents are mostly related to the product “orange juice”, or if those documents are for other products.



Results 1-10 of 84
(84/852 results matched)

You explored for "/"keyword$._word.noun.general"/"package" /"keyword$._word.noun.general"/"pack" -/"keyword$.product"/"lemo
language: English

Source	Title
Windows file system 4/10/08	orange juice - Package / container

... 1207826392281 Package / container Leak (bottom) orange juice When I was holding an unopened pack, the contents ... the bottom of the juice pack.

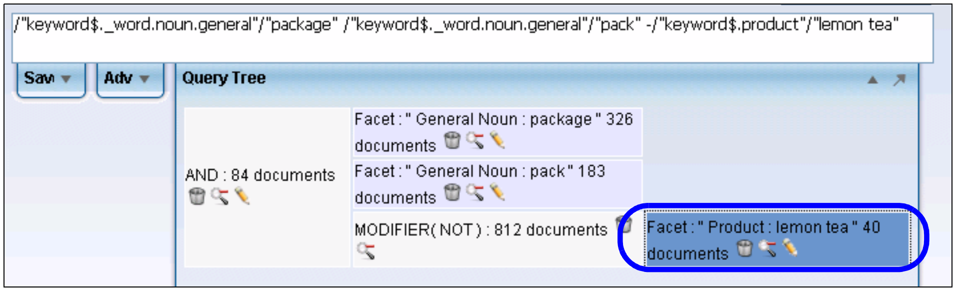
Source	Title
Windows file system 10/30/08	orange juice - Package / container

... 1225365592296 Package / container Leak (top) orange juice When I was holding an unopened pack,

Figure 5-46 The query result in the Documents view

To confirm how many documents are returned if the product is neither "lemon tea" nor "orange juice", follow these steps:

1. Select the **Product: lemon tea** node in the query tree (Figure 5-47).



/"keyword$._word.noun.general"/"package" /"keyword$._word.noun.general"/"pack" -/"keyword$.product"/"lemon tea"

AND : 84 documents

Facet : " General Noun : package " 326 documents

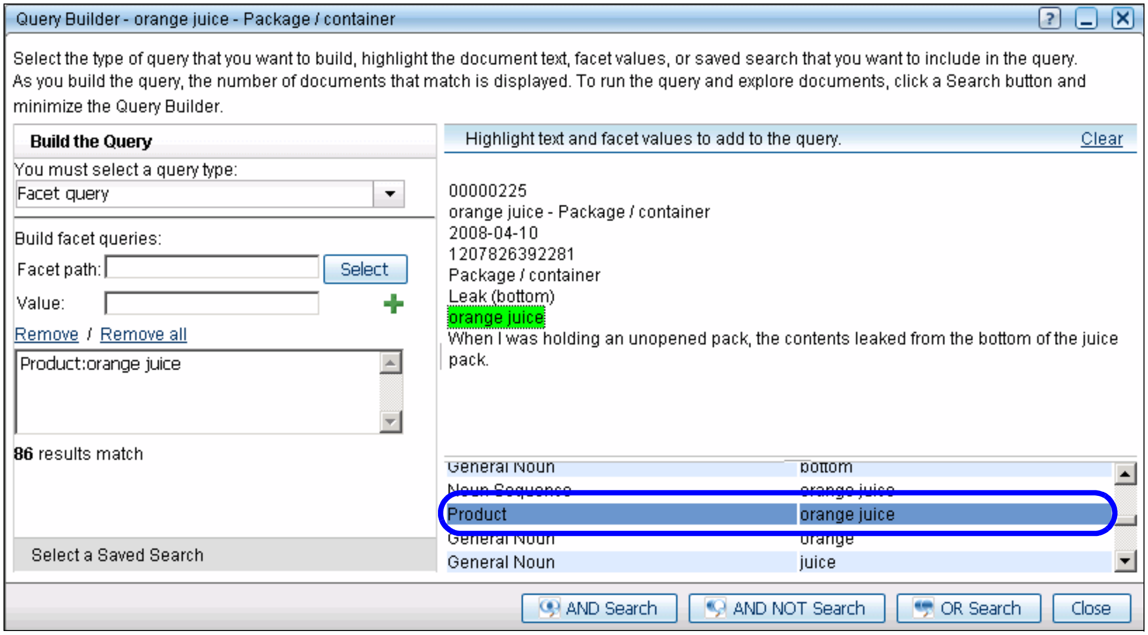
Facet : " General Noun : pack " 183 documents

MODIFIER(NOT) : 812 documents

Facet : " Product : lemon tea " 40 documents

Figure 5-47 Selecting a specific node in the query tree

2. Close the query builder for the first document (the document related to the "lemon tea" product), and open the query builder for the first document in the search results. This document mentions the "orange juice" product.
3. In the Query Builder window, select **Facet query** for the query type, and select the keyword **Product: orange juice** from the Facet list, as shown in Figure 5-48.



Query Builder - orange juice - Package / container

Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

Build the Query

You must select a query type: Facet query

Build facet queries:

Facet path:

Value:

Remove / Remove all

Product:orange juice

86 results match

Select a Saved Search

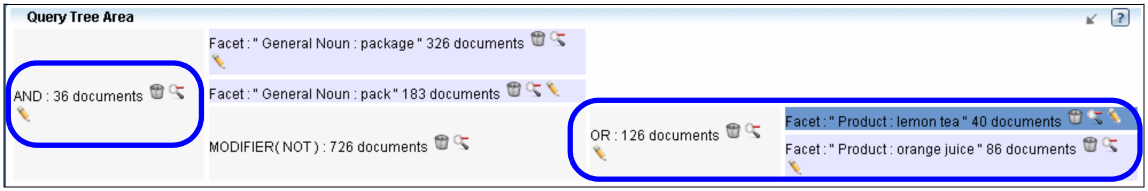
Highlight text and facet values to add to the query. Clear

00000225
orange juice - Package / container
2008-04-10
1207826392281
Package / container
Leak (bottom)
orange juice
When I was holding an unopened pack, the contents leaked from the bottom of the juice pack.

General Noun	bottom
Noun Sequence	orange juice
Product	orange juice
General Noun	orange
General Noun	juice

Figure 5-48 Opening the Query Builder window for another document and selecting another keyword

4. Issue the query with the OR search operator to add the query to the selected node. As shown in Figure 5-49, 36 documents match the query and are returned in the query results.



Query Tree Area

AND : 36 documents

Facet : " General Noun : package " 326 documents

Facet : " General Noun : pack " 183 documents

MODIFIER(NOT) : 726 documents

OR : 126 documents

Facet : " Product : lemon tea " 40 documents

Facet : " Product : orange juice " 86 documents

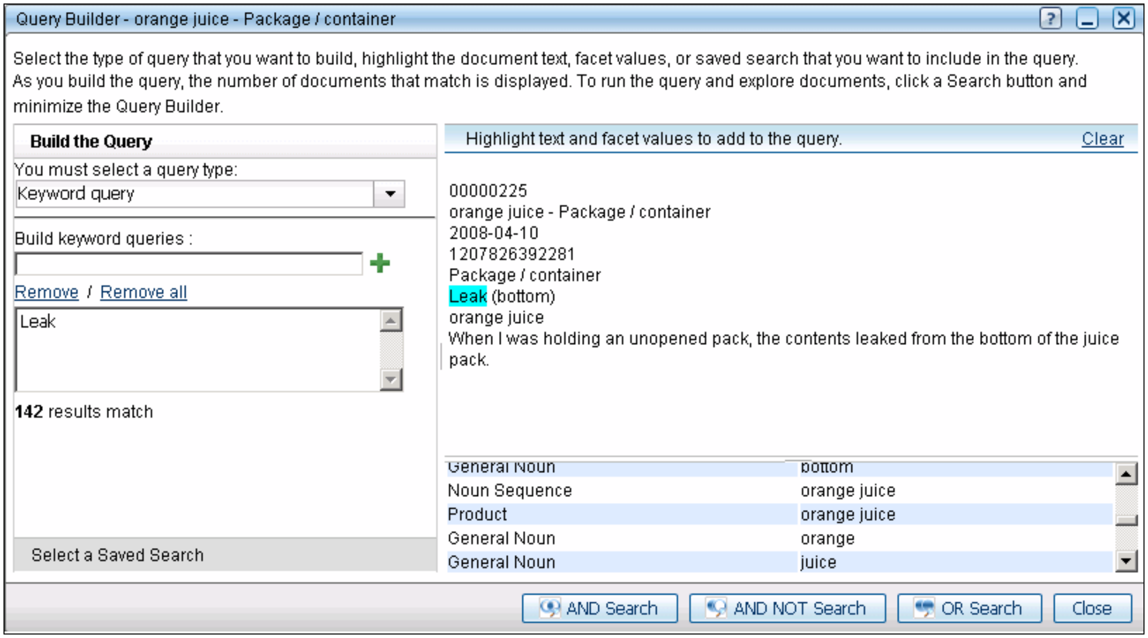
Figure 5-49   The results of using the OR search operator to add the query to the selected node

In our example, the query becomes “Product: lemon tea” keyword OR “Product: orange juice” keyword. However, we previously added the NOT operator before “Product: lemon tea.” Thus, the query returns 36 documents that represent documents that mention “package” and “pack,” but that are not for the products “lemon tea” nor “orange juice.”

Now that you have the result set, you can view the documents in the Documents view, or add other keywords to the query to gain insight into the content. When we look at the document in the Document view, we see the keyword “leak.”

To investigate further, follow these steps in the Query Builder window (Figure 5-50 on page 192):

1. Select the **Product: pack** node in the query tree.
2. Change the query type to **Keyword query**.
3. In the keyword field, type `leak`.
4. Click the **AND Search** button.



Select the type of query that you want to build, highlight the document text, facet values, or saved search that you want to include in the query. As you build the query, the number of documents that match is displayed. To run the query and explore documents, click a Search button and minimize the Query Builder.

You must select a query type: Keyword query

Build keyword queries :

Leak

142 results match

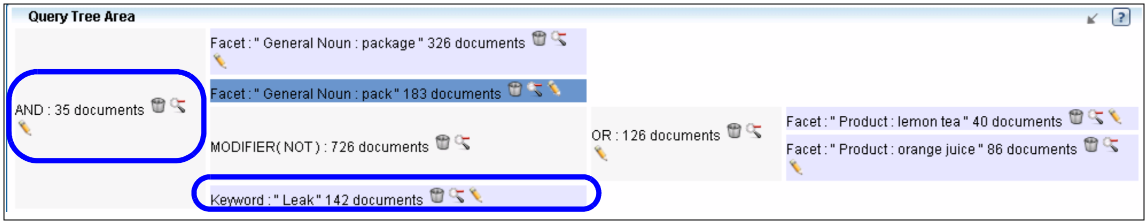
Highlight text and facet values to add to the query.

00000225
orange juice - Package / container
2008-04-10
1207826392281
Package / container
Leak (bottom)
orange juice
When I was holding an unopened pack, the contents leaked from the bottom of the juice pack.

General Noun	bottom
Noun Sequence	orange juice
Product	orange juice
General Noun	orange
General Noun	juice

Figure 5-50 Selecting another keyword as a keyword query

As a result, the query returns 35 documents, as shown in Figure 5-51.



AND : 35 documents

Facet : " General Noun : package " 326 documents

Facet : " General Noun : pack " 183 documents

MODIFIER(NOT) : 726 documents

OR : 126 documents

Facet : " Product : lemon tea " 40 documents

Facet : " Product : orange juice " 86 documents

Keyword : " Leak " 142 documents

Figure 5-51 The query tree for the query showing another keyword added with the AND operator

Based on this insight, the results show that 35 of the 36 documents results are related to the term “leak.”

To view the only document, based on the query, that does not mention the term leak, click the **Exclude** icon at the node “leak” in the query tree, as shown in Figure 5-52. As a result, the only document that is displayed (Figure 5-52) mentions “pack” and “package,” does not mention “leak,” and is not “Product: orange juice” nor “Product: lemon tea.”

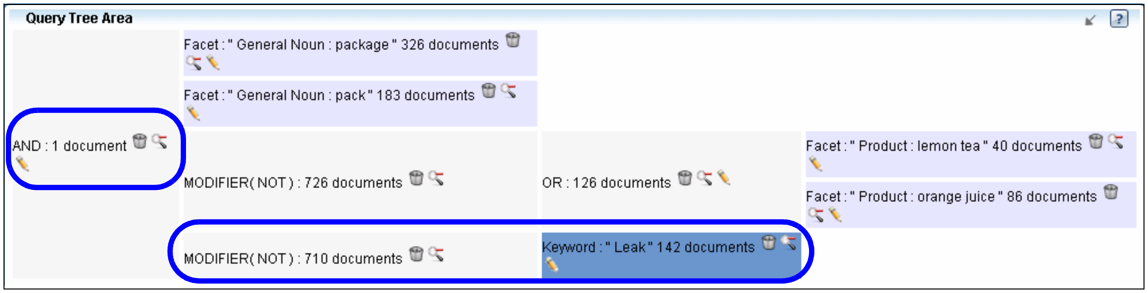

Query Tree Area
Facet : " General Noun : package " 326 documents
Facet : " General Noun : pack " 183 documents
AND : 1 document
MODIFIER(NOT) : 726 documents
OR : 126 documents
Facet : " Product : lemon tea " 40 documents
Facet : " Product : orange juice " 86 documents
MODIFIER(NOT) : 710 documents
Keyword : " Leak " 142 documents

Figure 5-52 The query tree after clicking the Exclude icon of the “leak” node

To review the content of the document in detail, open the Documents view.

5.4.4 Preferred practice for using the query builder and the query tree

As shown in 5.4.3, “Using the query builder” on page 186, you can interactively select a keyword used in the document by using the query builder and validating the query by using the query tree. One reason to use the query builder is that you can add the keyword to the query easily from the document preview. You can use the query builder to select a keyword from a document in the result set of the current built query or from the initial selected document. In addition, you can add a query to a selected node in the query tree.

The query builder and query tree aid with building a complex query so that you can go back and forth between the query builder and the query tree.

5.5 Rule-based categories with a query

When you build a query to find insight, you might want to use the query that you built to categorize the documents consistently. With rule-based categories, you can define a category and view documents by using a facet. This section explains how to configure the rule-based categories and how to use the rule-based categories.

5.5.1 Enabling the rule-based categories feature

To use the rule-based categories feature, you must enable the rule-based categorization type. Rule-based categorization is enabled for the collection by default.

Document clusters: To use document clustering, select **Rule-based and Document clusters** as the categorization type of the collection. For further information about document clustering, see 8.3, “Document clustering” on page 343.

In addition, if you want to add the query that you built from the text miner application, you must enable both the “Add rules to categories” and “Rebuild the category index” application user roles. By default, these user privileges are not enabled. Therefore, you must explicitly enable the roles by using the administration console.

Configuring the application user role: See “Configuring application user roles” on page 642 for further details about configuring the application user roles.

After you enable the “Add rules to categories” feature, an icon is displayed in the text miner application toolbar (Figure 5-53).

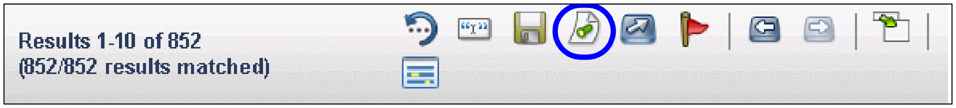

Results 1-10 of 852
(852/852 results matched)

Figure 5-53 Adding the current rule as a new category rule icon

5.5.2 Configuring rules for rule-based categories

With Content Analytics, you can define a category based on Uniform Resource Identifier (URI) patterns or Document content rules. You can use the query that you build to define the Document content rules and easily add a category rule from the text miner application. For further detail, go to the IBM Content Analytics Information Center at the following address, and search on *rule-based categories*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

5.5.3 Configuring rule-based categories

You can configure the category tree from scratch in the administration console. You can also modify the existing rule-based categories or delete categories. To access and configure a category tree, follow these steps:

1. From the administration console, go to the Collections view. In the list of collections, locate the collection that you want to edit, and click the **Edit** icon (Figure 5-54).

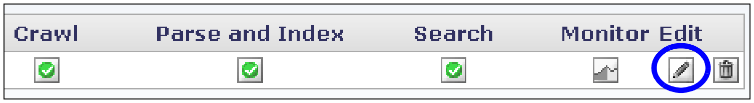
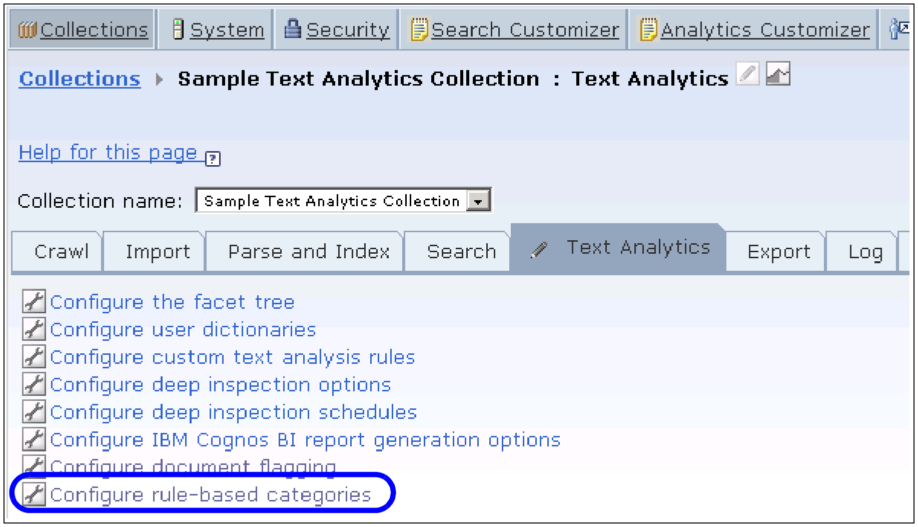


Figure 5-54 Editing and monitoring options for the collection

2. Select the **Text Analytics** tab and click **Configure rule-based categories** (Figure 5-55).

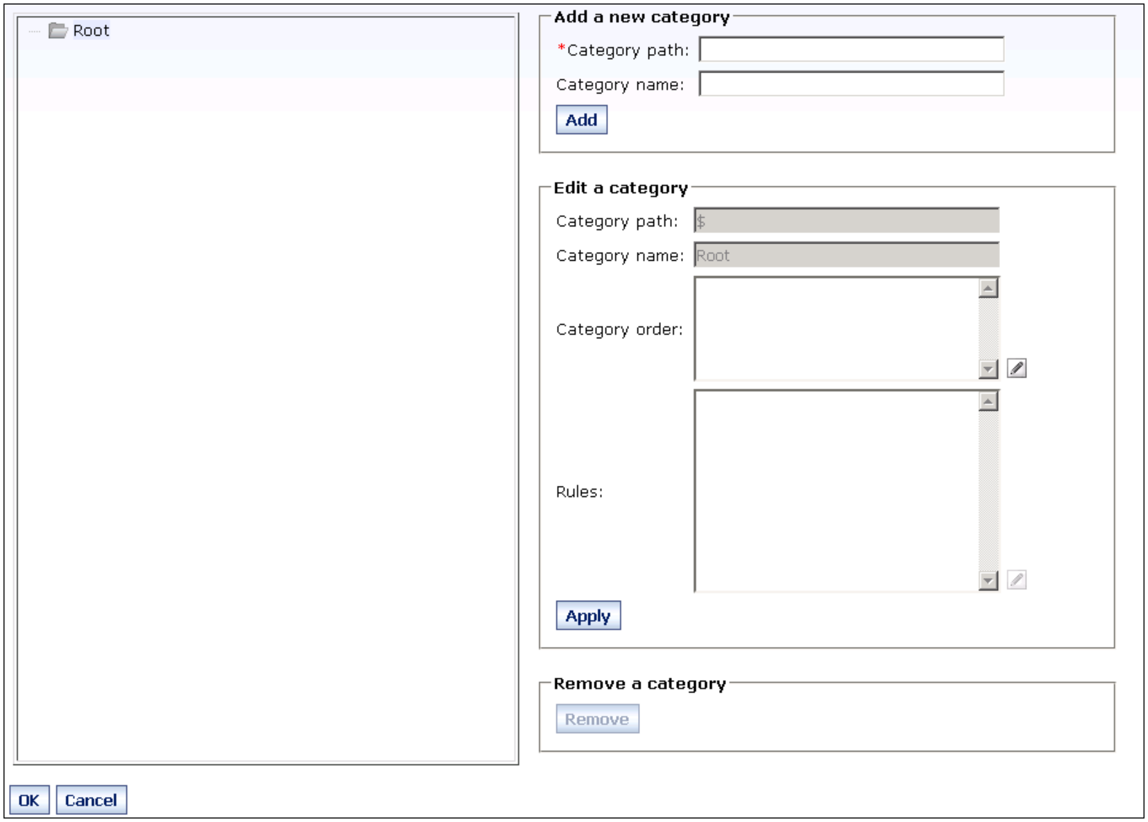

Collections
System
Security
Search Customizer
Analytics Customizer
Collections
Sample Text Analytics Collection : Text Analytics
Help for this page
Crawl
Import
Parse and Index
Search
Text Analytics
Export
Log
Configure the facet tree
Configure user dictionaries
Configure custom text analysis rules
Configure deep inspection options
Configure deep inspection schedules
Configure IBM Cognos BI report generation options
Configure document flagging
Configure rule-based categories

Collection name: Sample Text Analytics Collection

Figure 5-55 Text Analytics tab

> **Accessing the category tree:** Another way to access the category tree configuration page is to click the **Parse and Index** tab and, under the Select a categorization type section, select **Configure rule-based categories**.

The category tree is now displayed (Figure 5-56).

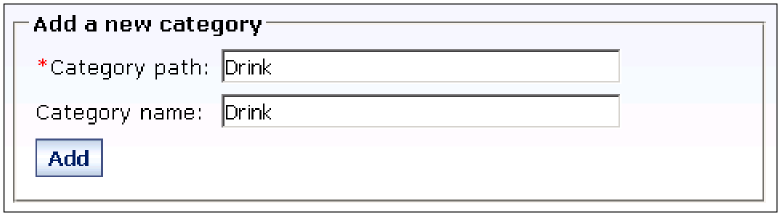


*Category path:
Category name:

Category path: $
Category name: Root
Category order:
Rules:

Figure 5-56 Category tree view in the administration console

3. To add a category, in the Add a new category section (Figure 5-57), complete the category path and name, and click **Add**.

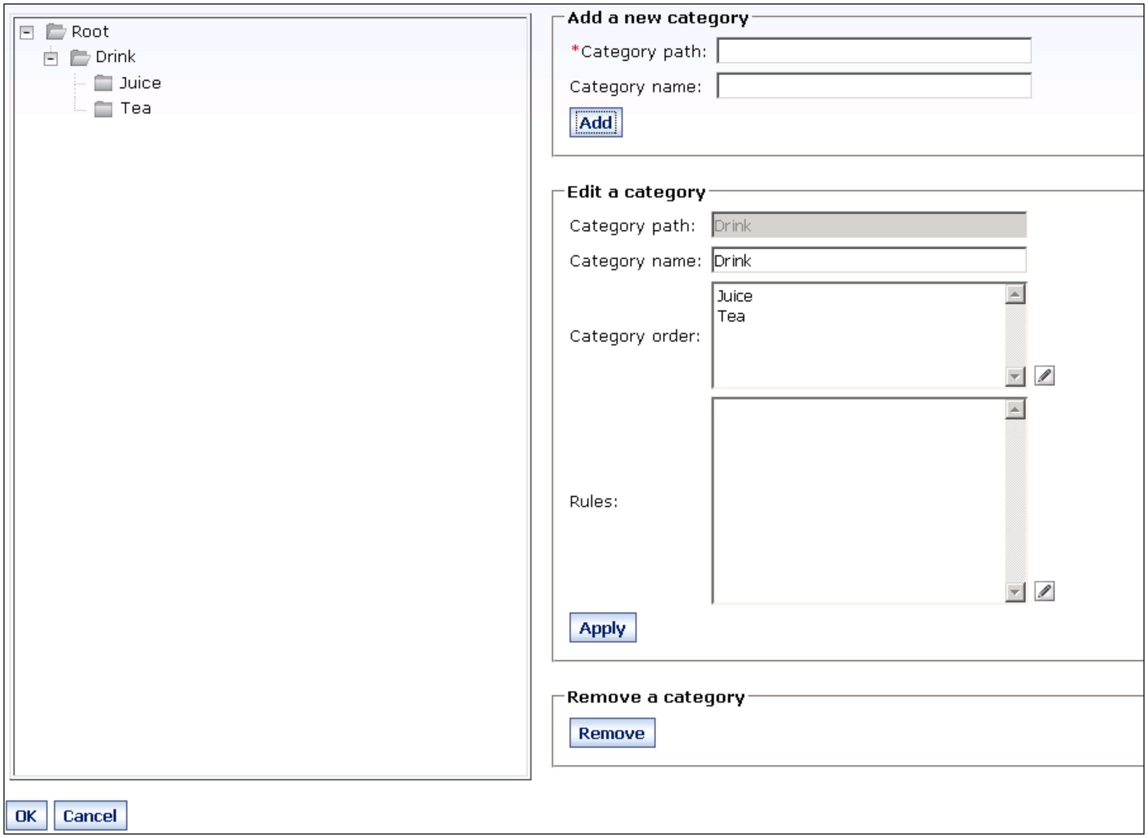


*Category path: Drink
Category name: Drink

Figure 5-57 Adding a category

After you add a category, you see the category in the category tree area (Figure 5-58).

4. To edit or delete the category, select the category in the category tree and perform the desired action. In our example, we define the Juice and Tea subcategories under the Drink category. The rule for each category is not defined at this point.



Category path: Drink

Category name: Drink

Category order: Juice
Tea

Rules:

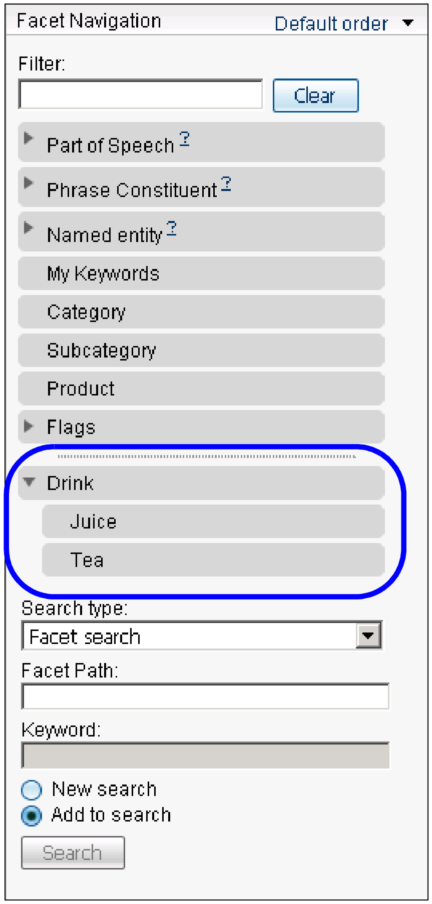
Apply

Figure 5-58 Adding, editing, or removing the category in the category tree

5. After you finish configuring the category tree, click **OK**.

> **Applying the category changes:** For category changes to take effect, restart the parse and index component, and redeploy the resources. Additionally, if the documents are already indexed by the component, rebuild the index after restarting the parse and index component.

After you define the new categories and rebuild the index, the defined category is shown as a new facet in the Facet Navigation pane (Figure 5-59). In our example, we added a Drink category with two subcategories, Juice and Tea. As a result, the Drink facet is shown with two children facets, Juice and Tea.



Search type: Facet search

○ New search
● Add to search

Figure 5-59 Facet Navigation after the rule-based category is added

5.5.4 Adding the current query as a category rule

As explained earlier, you can define a category from the administration console. To define a category rule, you must know the details of the rule to add it. You might not know your desired category rule until you start to analyze your content by using the text miner application.

With Content Analytics, you can define the category rule based on a query that you want to use for document categorization within the text miner application.

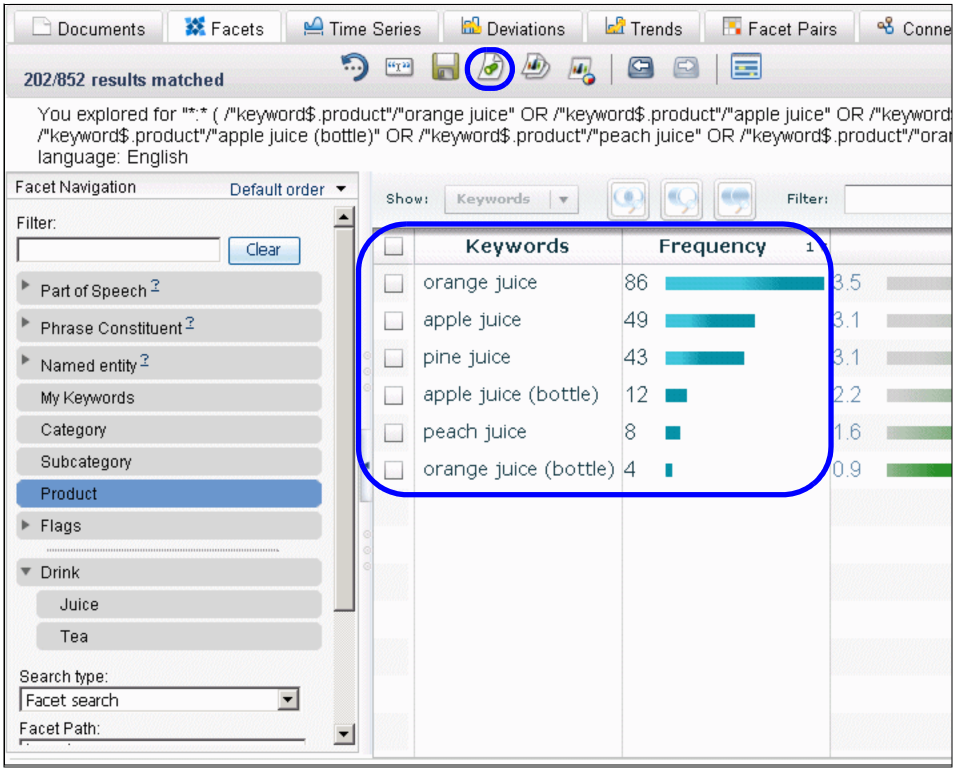
This feature of adding the query to the category rule within the text miner application is useful because it eliminates the need to go back to the administration console to define the rule.

For example, consider a case when you want to see the documents that are related to the term “juice.” You filter the product by the term “juice” in the Facets view, and add the filtered keywords to the query with the AND operator. Example 5-1 shows the resulting query for this example.

Example 5-1 The query filtered by the term “juice” in the Product facet

```
/"keyword$.product"/"orange juice" OR /"keyword$.product"/"apple juice"
OR /"keyword$.product"/"pine juice" OR /"keyword$.product"/"apple juice
(bottle)" OR /"keyword$.product"/"peach juice" OR
/"keyword$.product"/"orange juice (bottle)"
```

The query returns 202 documents, as shown in Figure 5-60.



202/852 results matched

You explored for "*:* (/"keyword$.product"/"orange juice" OR /"keyword$.product"/"apple juice" OR /"keyword
/"keyword$.product"/"apple juice (bottle)" OR /"keyword$.product"/"peach juice" OR /"keyword$.product"/"orai
language: English

Search type: Facet search

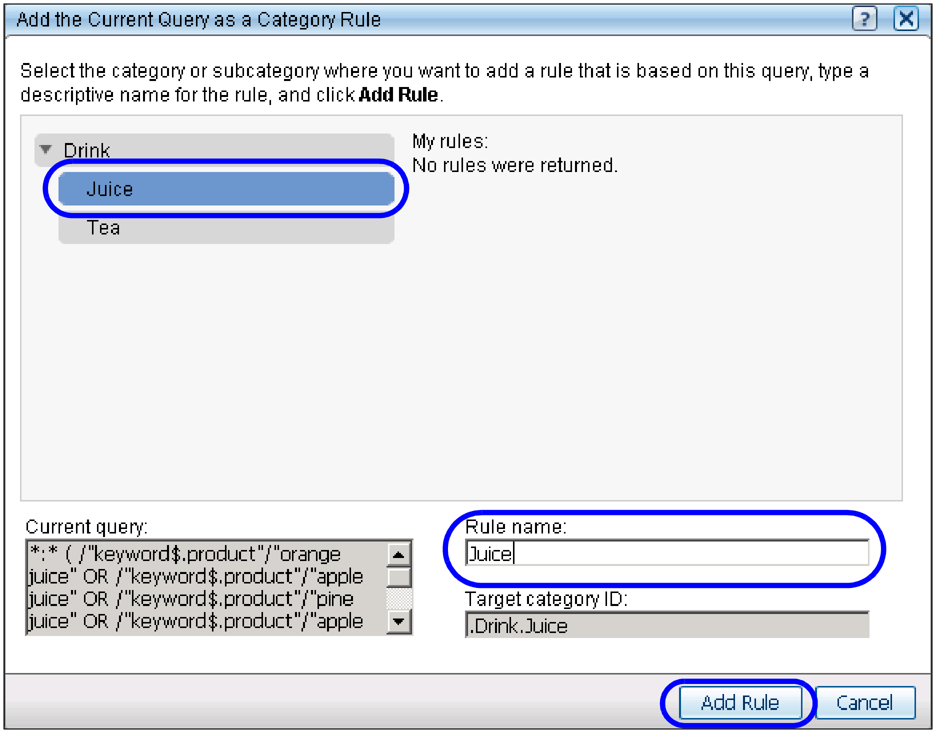
Facet Path:

□	Keywords	Frequency	1
□	orange juice	86	3.5
□	apple juice	49	3.1
□	pine juice	43	3.1
□	apple juice (bottle)	12	2.2
□	peach juice	8	1.6
□	orange juice (bottle)	4	0.9

Figure 5-60 The 202 documents returned by the query

To add the query as a category rule, follow these steps:

1. Click the **Add the current query as a category rule** icon (circled in Figure 5-60 on page 199).
2. In the Add the Current Query as a Category Rule window (Figure 5-61), complete these steps:
   a. Verify that the current query that is shown is correct.
   b. Select the category that you want to use. For our example, we click the **Juice** facet.
   c. Type the rule name. We type `Juice` in the Rule name field.
   d. Click **Add Rule** to add the current query as a category rule for the selected category.



Select the category or subcategory where you want to add a rule that is based on this query, type a descriptive name for the rule, and click **Add Rule**.

My rules:
No rules were returned.

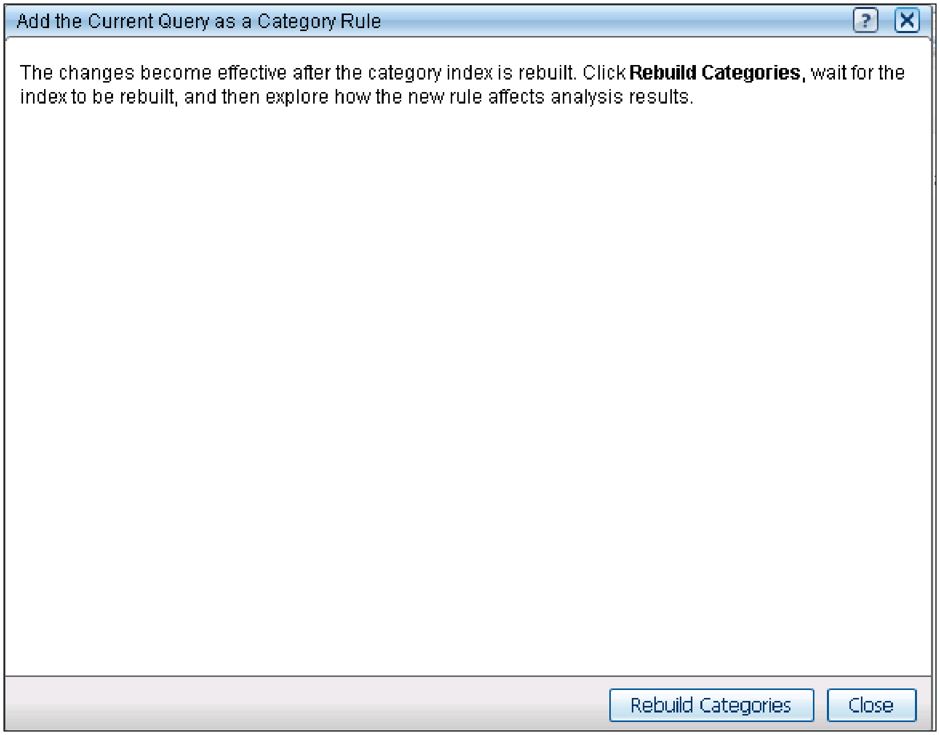
Current query:
: (/"keyword$.product"/"orange juice" OR /"keyword$.product"/"apple juice" OR /"keyword$.product"/"pine juice" OR /"keyword$.product"/"apple

Rule name: Juice

Target category ID: .Drink.Juice

Figure 5-61 Adding the current query as a category rule

3. In the message window that opens (Figure 5-62), click **Rebuild Categories** to start the document categorizer to rebuild the categories.

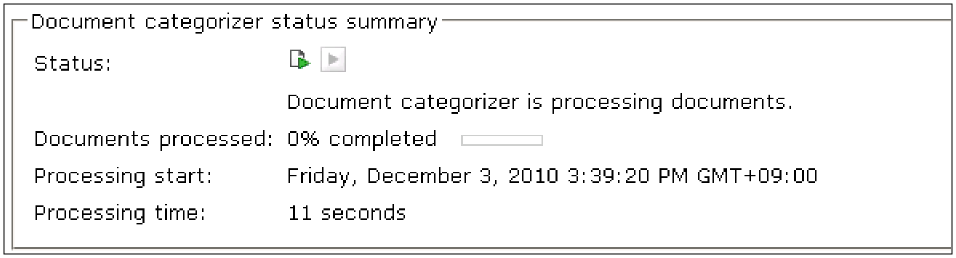


The changes become effective after the category index is rebuilt. Click **Rebuild Categories**, wait for the index to be rebuilt, and then explore how the new rule affects analysis results.

Figure 5-62 Starting to rebuild the categories after adding the query as a category rule

Rebuilding the categories: Make sure the parse and index process of the collection is up and running.

To monitor the progress of the document categorizer from the administration console, click the **Parse and Index** tab for your collection, and click **Details**. Figure 5-63 shows the status of the progress.


Document categorizer status summary

Status:

Document categorizer is processing documents.

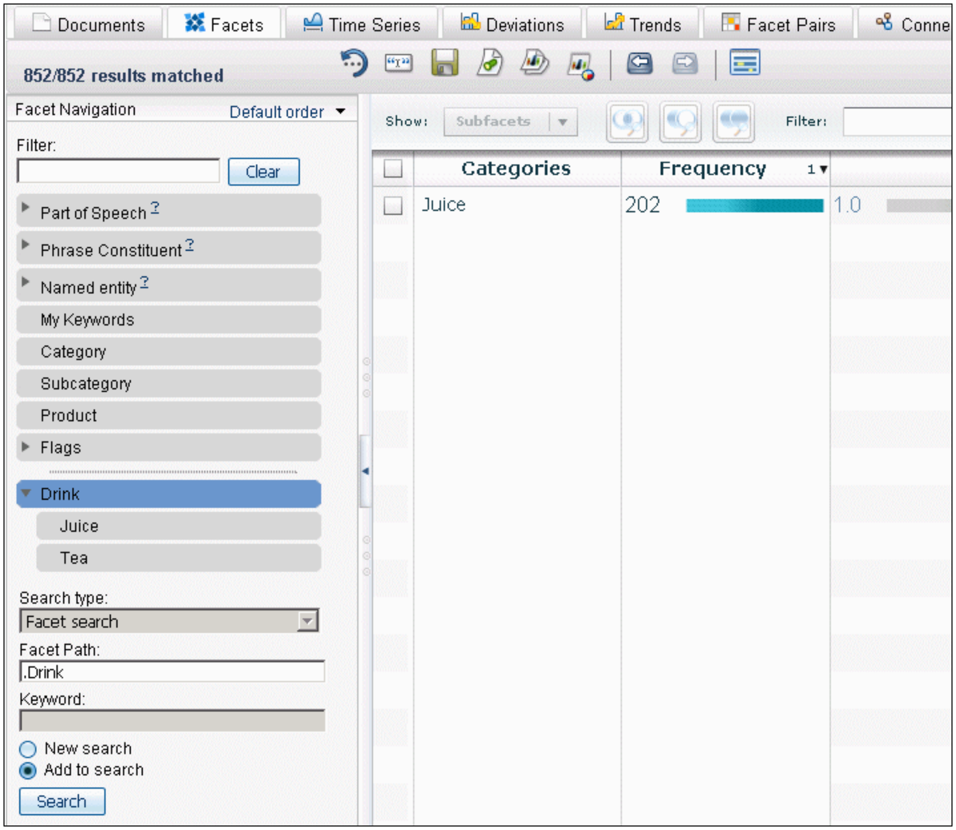
Documents processed: 0% completed

Processing start: Friday, December 3, 2010 3:39:20 PM GMT+09:00

Processing time: 11 seconds

Figure 5-63 Checking the Document categorize status on the administration console

After the document categorizer process is completed, the new category facets are displayed in the text miner application (Figure 5-64).



852/852 results matched

Filter:

Search type: Facet search

Facet Path: .Drink

Keyword:

○ New search
● Add to search

□	Categories	Frequency 1 ▾	
□	Juice	202	1.0

Figure 5-64 New categories returned

In our example, we added the query shown in Example 5-1 on page 199 as a category rule to the Juice category. The result is that 202 documents make up the Juice category. This result set is the same as searching for the query described in Example 5-1 on page 199. However, the column label is different for the two situations. The column label is “Keywords” for the direct query search (Figure 5-60 on page 199), and it is “Categories” when using rule-based categories (Figure 5-64 on page 202).

You can add the query as a category rule and use the category to find further insight into a set of documents. To build a query, you can use the query tree or query builder functions to interactively perform your analysis.

5.6 Common view features

Content Analytics provides several features that are common to all views in the text miner application. These features are on the menu bar under the view tabs. Except for the Search result counts label, each feature is identified and activated by its own icon.

See Table 5-3 for an explanation of each feature.

Table 5-3 Common label and icons in the analysis window

Label and icon	Description
Search result counts 46/852 results matched	At all times, Content Analytics shows the number of documents that match the current query out of the total number of documents in the entire corpus of documents. These statistics are for your reference.
Clear the current condition	Click this icon to remove the current query condition from the search box, thus resetting the documents that are being analyzed back to the entire corpus of documents. Use this function when you want to start a new analysis with a different query condition. Because the current query condition is not saved automatically, you might want to click the **Save the search** button before you click the **Clear the current condition** button. This operation is the same operation when you click the **Clear** button in the search field.
Show and hide query input area	Click this icon to show or hide the search field area. This icon works the same as when you click the **Show query input area** link. After you click this icon, the search box is either displayed or hidden at the top of the window.
Save the search	Click this icon to save your query condition. A window opens that prompts you to enter the name of your query. You can also provide a description for the query. After you save the query, you can retrieve the query by its name from the **Saved Searches** tab.

Label and icon	Description
Export the results	Click this icon to export your search results. This button is displayed only if exporting searched documents is enabled.
Go Back/Forward to a query to rerun	These icons behave similar to the back and forward buttons of a browser. Click these icons to move back or forward through each query progression when it is built during your analysis. Click the icon to execute the previous or next query and refresh the search results of all views. This capability applies to the current browser session. Information is not preserved permanently even if you log in to the application when the security is enabled. Also, when you click the **Clear the current condition** button, the history is cleared.
Show/Hide document properties	Click these icons to show or hide the detailed document properties, such as the DocumentID and Title in the Documents view, for each document.
Specify preferences for viewing search results	When you click this icon, the same window opens as when you click the **Preferences** link in the application toolbar. When you click this icon, the preference for the view that you are using or results tab is displayed. See 5.1.4, “Changing the default behavior by using preferences” on page 150, for a description of what you can set in this window. The **Results** tab is selected by default when you click this icon.
Add the current query as a new category rule	When you click this icon, a window opens in which you can add the current query as a new category rule for an existing category. This button is displayed only if the rule-based category is enabled. For further information about rule-based categories, see 5.5, “Rule-based categories with a query” on page 193.
Set and clear Document flags	This button is displayed only in the Documents view. It is displayed if document flagging is enabled and the “Manage document flags” application user role is enabled. For further information, see 5.7, “Document flagging” on page 205.
Create deep inspection reports	When you click this icon (except in the Documents view), you can create a deep inspection report. You must select a facet to create a deep inspection report. This button is displayed when the “Create deep inspection reports” application user role is enabled. For further information, see 10.7, “Deep inspection” on page 431.
Create a report for Cognos BI or download the report as CSV file	When you click this icon (except in the Documents view), a window opens. You specify whether you want to save the output in the comma-separated values (CSV) format and where to save the output, or you can specify how to create the Cognos BI report. This button is displayed if “Create IBM Cognos BI reports” is enabled for the collection. For further information, see Chapter 13, “Integrating Cognos Business Intelligence” on page 525.

Icons for enabled features: Some of the icons in Table 5-3 on page 203 are displayed only if you enabled that particular feature.

5.7 Document flagging

With document flagging, you can assign a custom flag to a single document or a group of documents for classification, export, or additional analysis purposes. This feature is convenient after you perform multiple searches to find the set of documents that you want to further examine, to export them, or to classify them. The administrator defines the document flags that are selectable by users in the text miner application. After the documents are assigned with a flag, users can narrow down the documents by using the flag or view the flag count on the document result set for further analysis.

5.7.1 Configuring document flags

Configure document flags that will be selectable within the text miner application. In this scenario, you create two document flags named Public Relations and Quality Assurance.

Flags in a collection: A collection can contain update to 64 document flags.

To configure document flags, perform these steps:

1. Open the administration console, and click the **Edit** icon for the Sample Text Analytics Collection.
2. Click the **Text Analytics** tab and click the **Configure document flagging** link (Figure 5-65).

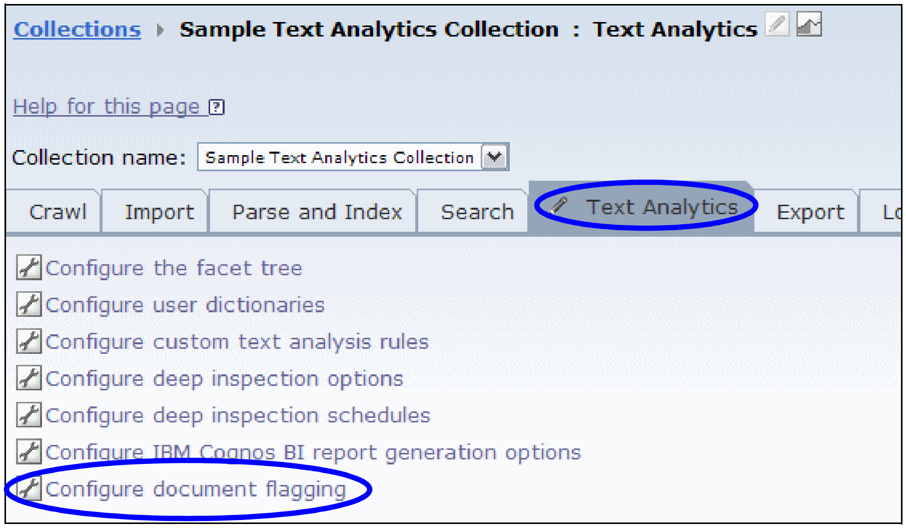
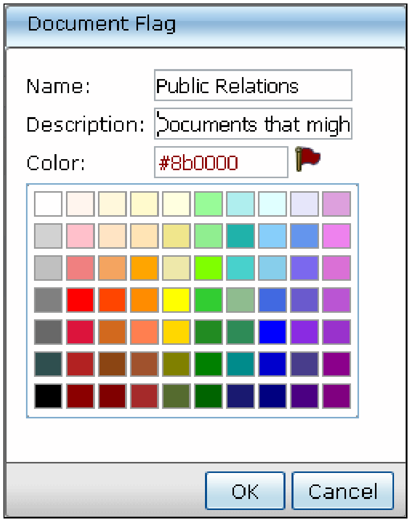

Collections
Sample Text Analytics Collection : Text Analytics
Help for this page
Collection name:
Sample Text Analytics Collection
Crawl
Import
Parse and Index
Search
Text Analytics
Export
Lo
Configure the facet tree
Configure user dictionaries
Configure custom text analysis rules
Configure deep inspection options
Configure deep inspection schedules
Configure IBM Cognos BI report generation options
Configure document flagging

Figure 5-65 Configuring document flagging

3. In the Document Flag window (Figure 5-66), complete these steps:
   a. Click **Add New Document Flag**.
   b. In the Name field, type `Public Relations`.
   c. In the Description field, type `Documents that might result in public relation exposure`.
   d. In the Color field, select a red color or type `#8b0000`.
   e. Click **OK** to add the new document flag.



Name: Public Relations

Description: Documents that migh

Color: #8b0000

Figure 5-66 Adding a document flag

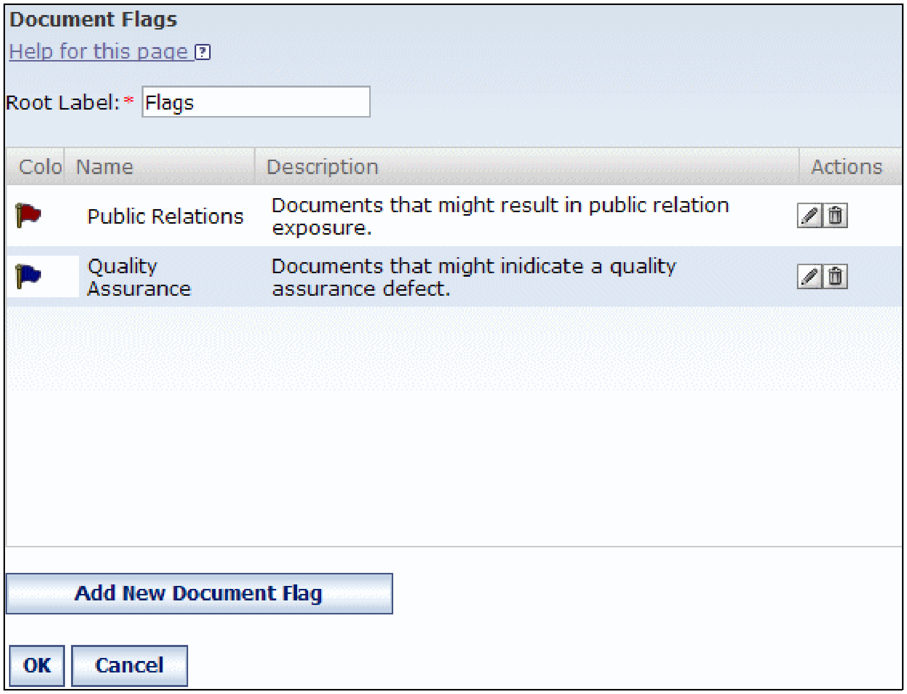
4. Repeat step 3 but use the field values indicated in Table 5-4.

Table 5-4 Document flag properties

Field	Value
Name	Quality Assurance
Description	Documents that might indicate a quality assurance defect.
Color	Blue or #000080

The document flag window now looks similar to the example shown in Figure 5-67. As a result, the Public Relations and Quality Assurance facets are displayed under the Flags facet in the text miner application.

5. Click **OK**.



Document Flags

Help for this page

Root Label:* Flags

Colo	Name	Description	Actions
	Public Relations	Documents that might result in public relation exposure.	
	Quality Assurance	Documents that might inidicate a quality assurance defect.	

Figure 5-67 List of defined document flags

Authority to manage document flags: Users must have proper authority to manage document flags in order to add and remove flags associated with documents. If users are required to log in to the text miner application, select **Security** → **Application User Roles** to provide the user with authority to manage document flags. If users are not required to log in to the text miner application, select **Security** → **System Security** → **Configure application user roles** to provide the authority.

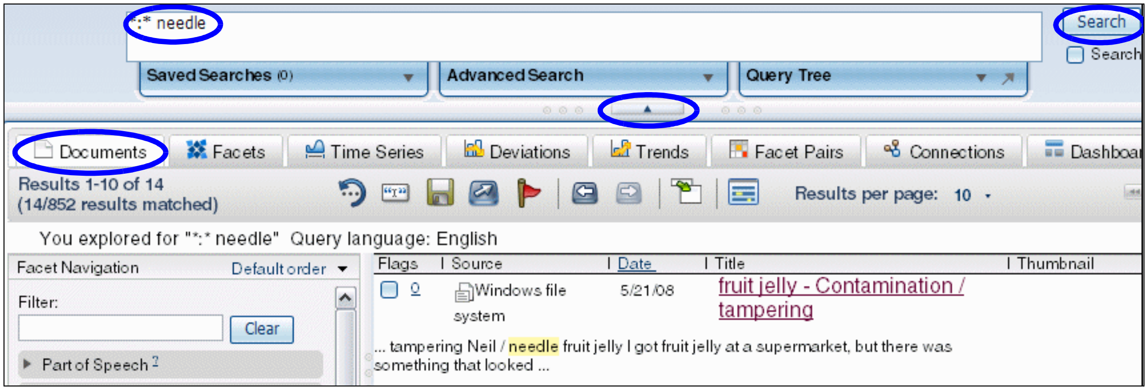
5.7.2 Setting document flags

Now that you configured the document flags, you can select and view them in the text miner application. For our scenario, a new root facet named Flags, with Public Relations and Quality Assurance facets as children, is displayed in the text miner facet window. This section explains how to set these new document flags.

Associating all search documents with a document flag

To associate all search documents with a document flag, follow these steps:

1. Open the text miner application.
2. Expand the query text area, and type `needle`. Click **Search** to search for the term `needle` (Figure 5-68). Now the document result set contains documents that are related to needle.



Results 1-10 of 14
(14/852 results matched)

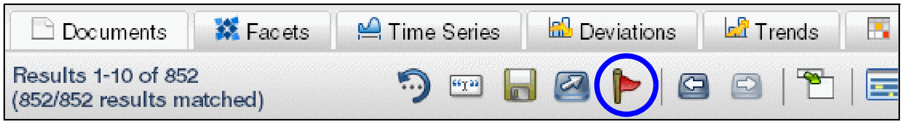
You explored for "*:* needle" Query language: English

Flags	Source	Date	Title	Thumbnail
□	Windows file system	5/21/08	fruit jelly - Contamination / tampering	

... tampering Neil / needle fruit jelly I got fruit jelly at a supermarket, but there was something that looked ...

Figure 5-68 Documents containing the term "needle"

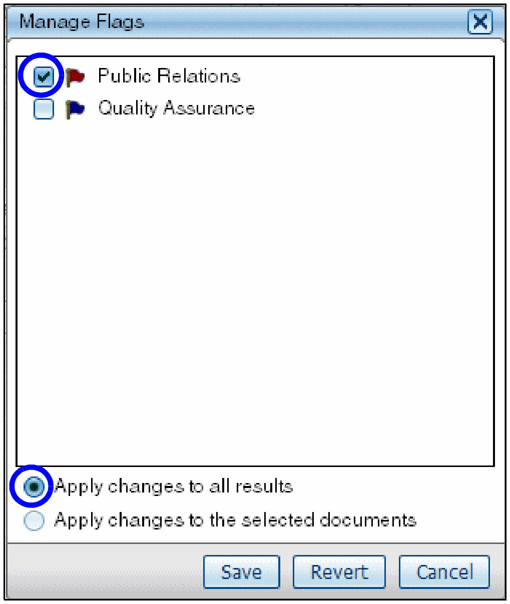
3. Click the **Documents** tab.
4. Click the **Document Flag** icon (highlighted in Figure 5-69).



Results 1-10 of 852
(852/852 results matched)

Figure 5-69 Document Flag icon in the text miner application

5. In the Manage Flags window (Figure 5-70), select **Public Relations**, and click **Allow changes to all results**. This action marks every document in the query result set with a Public Relations flag. Click **Save**.



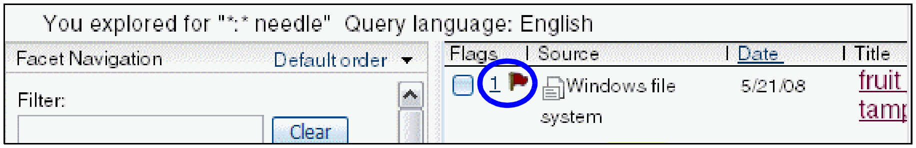
Manage Flags

☑ Public Relations
□ Quality Assurance

● Apply changes to all results
○ Apply changes to the selected documents

Figure 5-70 Selecting document flags for query results

Notice that the documents now have a red flag and the number 1 next to them in the Document view table, as shown in Figure 5-71. The number indicates how many document flags are associated with this document. In this scenario, one document flag is associated so far.



You explored for "*:* needle" Query language: English

Flags	Source	Date	Title
□ 1	Windows file system	5/21/08	fruit tamp

Figure 5-71 Document associated with the Public Relations flag

Adding a document flag to a single document with quick flags

You can add another document flag to a document after you have already assigned a document flag. With quick flags, you can quickly associate a flag to a document without selecting the document.

1. In the Documents view, click the **1** flag link for the first document listed in the table (Figure 5-71 on page 210).
2. In the quick flag pop-up window, select the **Quality Assurance** flag (Figure 5-72). This action automatically adds the Quality Assurance flag to the document you selected.

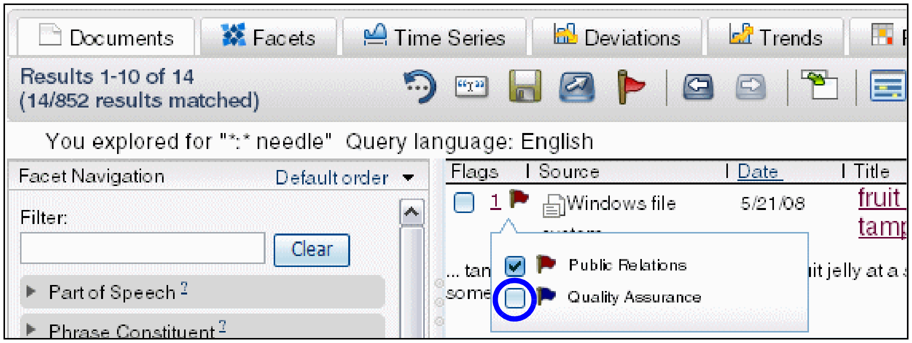


Figure 5-72 Adding a document flag to a single document using the quick flag link

Now the document shows two flag icons associated with it, and it contains a 2 link to indicate two document flags, as shown in Figure 5-73. One of the flag icons is the color of the Public Relations flag, and the other flag is the color of the Quality Assurance flag.

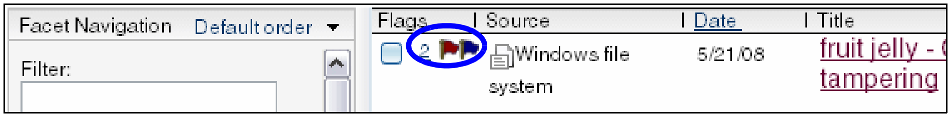
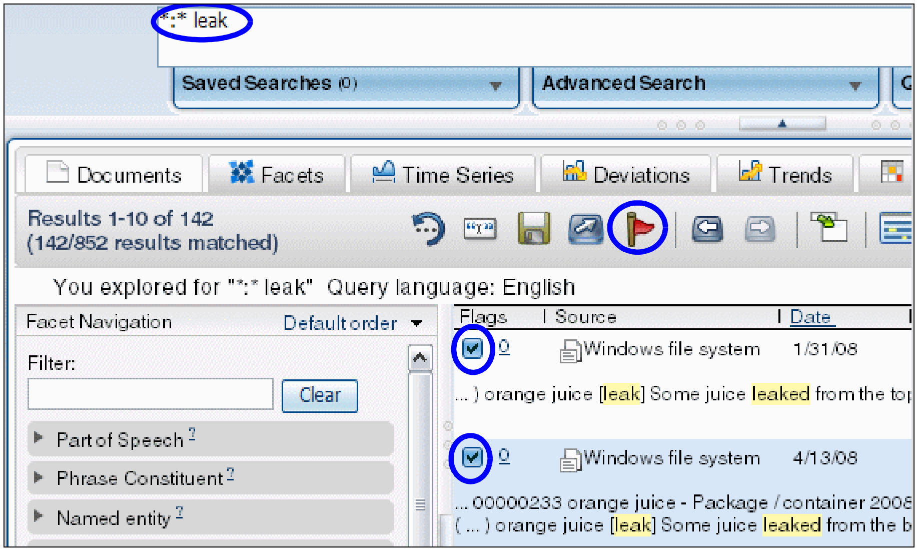


Figure 5-73 Document with two flags associated to it

Associating selected documents to a document flag

You can associate a selected set of documents to a particular flag rather than associating the flag to the entire query result set:

1. Click **Clear** in the query text area to clear the search query to start a new search view.
2. Click the **Documents** tab (Figure 5-74).
3. In the query search text area, type `leak`. This action shows all documents that contain the term “leak” so that you can analyze them further.
4. Select the first two documents in the result set by clicking the check box to the left of the document.



: leak

Results 1-10 of 142
(142/852 results matched)

You explored for "*:* leak" Query language: English

Flags	Source	Date
☑ 0	Windows file system	1/31/08
☑ 0	Windows file system	4/13/08

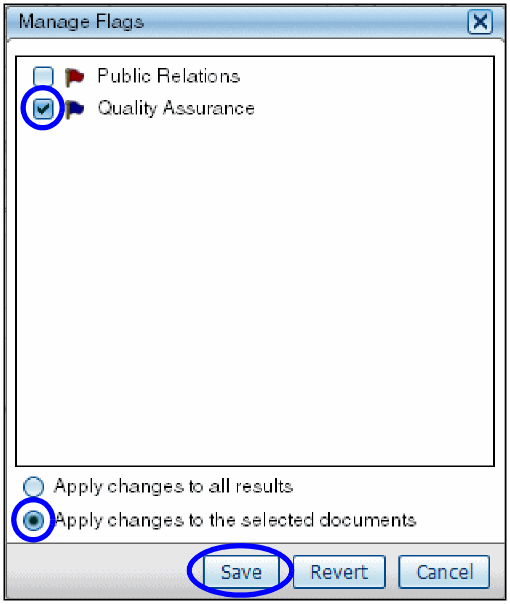
...) orange juice [leak] Some juice leaked from the to

... 00000233 orange juice - Package / container 2008
(...) orange juice [leak] Some juice leaked from the b

Figure 5-74 Selecting documents to assign the document flag

5. Click the **Flags** icon (highlighted in Figure 5-74).

6. In the Manage Flags window (Figure 5-75), select the **Quality Assurance** check box, and click **Apply changes to the selected documents**. Then click **Save**.


Manage Flags

☐ Public Relations
☑ Quality Assurance

○ Apply changes to all results
● Apply changes to the selected documents

Figure 5-75 Applying the flag changes to selected documents

5.7.3 Viewing the document values of a flag facet

With the facet view, you can view all of the documents that are associated with one flag. You can narrow down the documents by flag or view the flag count on the returned documents. To view the documents associated with the Public Relations document flag, follow these steps:

1. Click **Clear** to clear the search query to start a new search view.
2. Click the **Facet** tab.

3. Expand the **Flags** facet, and count the documents for each flag that is displayed (Figure 5-76).

4. To view all of the documents marked with the Public Relations flag, select the **Public Relations** facet check box and click the **Add to search with Boolean AND** icon (highlighted in Figure 5-76).

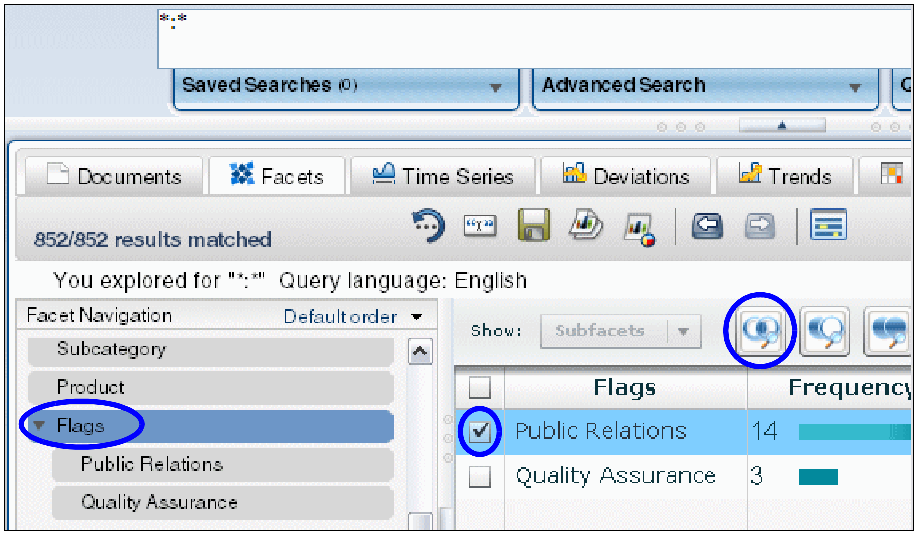
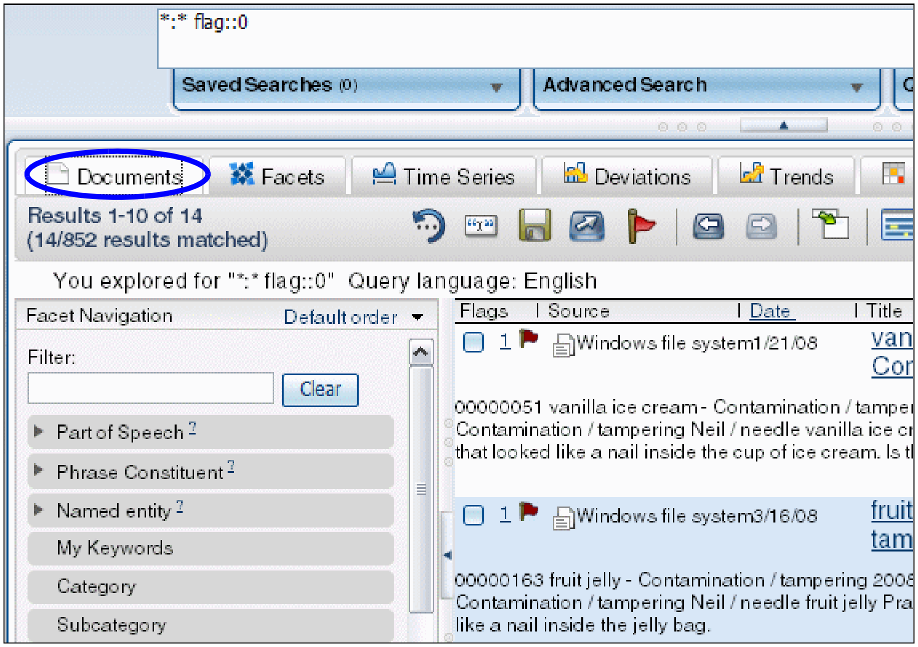


Figure 5-76   Display of document flag facets

5. Click the **Documents** tab (Figure 5-77) to view and further analyze all of the documents that are associated with the Public Relations flag.



: flag::0

Results 1-10 of 14
(14/852 results matched)

You explored for "*:* flag::0" Query language: English

Flags	Source	Date	Title
□ 1	Windows file system	1/21/08	van Cor
□ 1	Windows file system	3/16/08	fruit tam

00000051 vanilla ice cream - Contamination / tamper
Contamination / tampering Neil / needle vanilla ice cr
that looked like a nail inside the cup of ice cream. Is t

00000163 fruit jelly - Contamination / tampering 2008
Contamination / tampering Neil / needle fruit jelly Pra
like a nail inside the jelly bag.

Figure 5-77 Documents associated with the Public Relations facet

The text miner application provides many views for content analysis. For information about the text miner application views, see Chapter 6, “Text miner application: Views” on page 217. If you are already familiar with all the views, proceed with Chapter 7, “Performing content analysis” on page 279.

6

Text miner application: Views

Chapter 5, “Text miner application: Basic features” on page 143, provides information about the basic features of the text mining application, with a specific focus on the search and discovery features. This chapter focuses on the text miner application views, their features, and their functions.

Specifically, this chapter includes the following sections:

- Views
- Documents view
- Facets view
- Time Series view
- Trends view
- Deviations view
- Facet Pairs view
- Connections view
- Dashboard view

If you are already familiar with the user interface of the text miner application, including its views and search and discovery features, proceed to Chapter 7, “Performing content analysis” on page 279.

6.1 Views

The text miner application provides the following views to assist in text mining. These views (shown in Figure 6-1) are generated dynamically based on your query and facet selections.

Documents view Shows a list of documents that match your query.

Facets view Shows a list of keywords for a selected facet.

Time Series view Shows the frequency change over time.

Trends view Shows sharp and unexpected increases in frequency over time.

Deviations view Shows the deviation of keywords for a given time period.

Facet Pairs view Shows the correlation of keywords from two selected facets.

Connections view Shows the correlation of keywords from two selected facets in a graphical way.

Dashboard view Shows a configured dashboard layout with one or more graphs or tables in a single view.

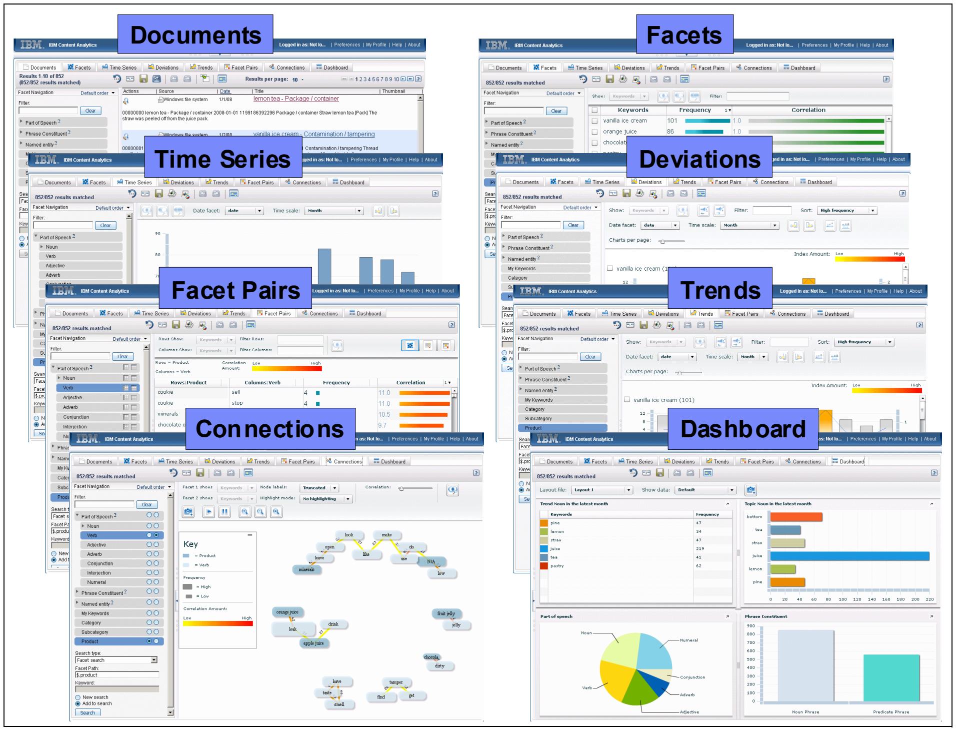


Figure 6-1 All views available in the text miner application

For information about the general user interface of the text miner application, see 5.1, “Overview of the text miner application” on page 144. For information about the search and discovery features of the text miner application, see 5.2, “Search and discovery features” on page 152.

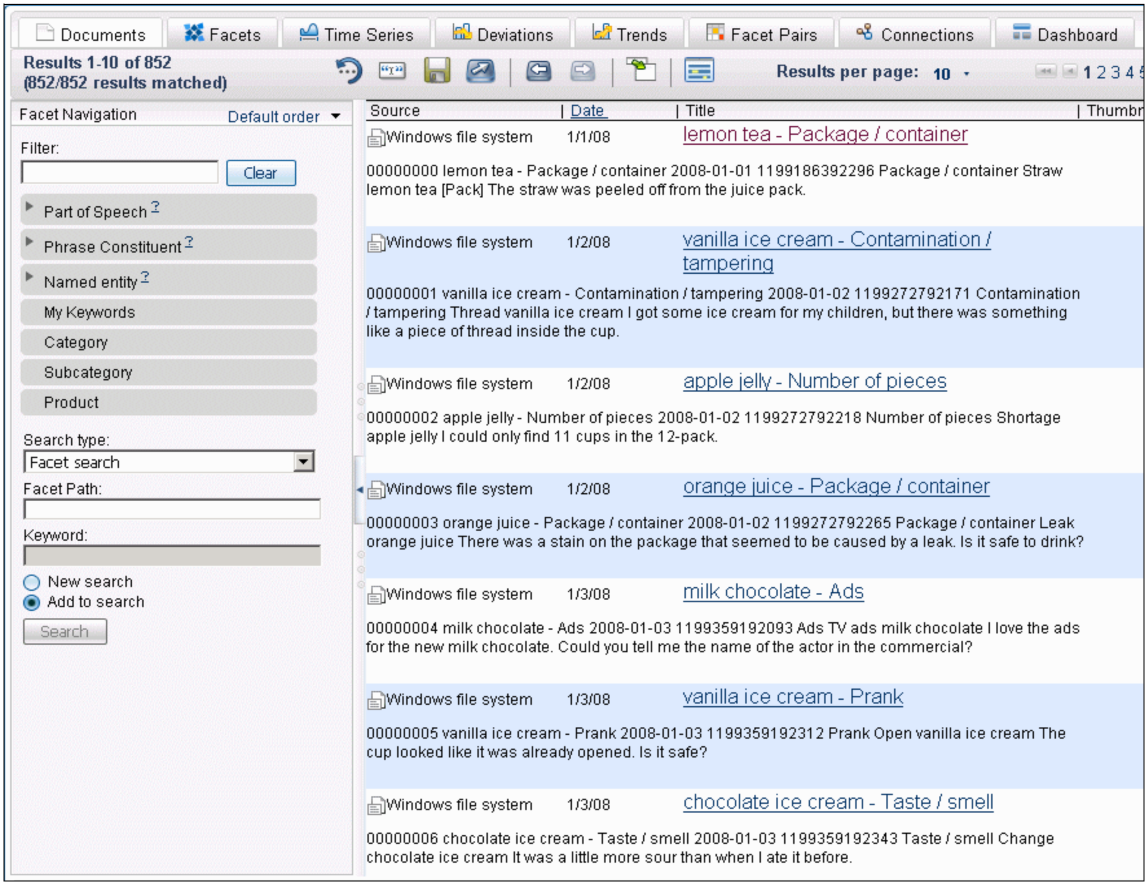
6.2 Documents view

The Documents view shows a list of documents that match your query or facet selections. You use this view when you want to see the contents of an individual document. By default, ten results are displayed per page in the Documents view. You can click the page buttons to navigate between results pages or to move to a specific page.

Each entry in the document list contains the following information:

- A dynamic summary of the document based on your query terms
- The source and date of the document
- A thumbnail of the document (if configured)
- A document link displayed as the title, which, when clicked, shows the original document from its crawled source

Figure 6-2 shows the Documents view with the sample documents. The default search condition is `*:*`.



Results 1-10 of 852
(852/852 results matched)

Source	Date	Title
Windows file system	1/1/08	lemon tea - Package / container

00000000 lemon tea - Package / container 2008-01-01 1199186392296 Package / container Straw lemon tea [Pack] The straw was peeled off from the juice pack.

Source	Date	Title
Windows file system	1/2/08	vanilla ice cream - Contamination / tampering

00000001 vanilla ice cream - Contamination / tampering 2008-01-02 1199272792171 Contamination / tampering Thread vanilla ice cream I got some ice cream for my children, but there was something like a piece of thread inside the cup.

Source	Date	Title
Windows file system	1/2/08	apple jelly - Number of pieces

00000002 apple jelly - Number of pieces 2008-01-02 1199272792218 Number of pieces Shortage apple jelly I could only find 11 cups in the 12-pack.

Source	Date	Title
Windows file system	1/2/08	orange juice - Package / container

00000003 orange juice - Package / container 2008-01-02 1199272792265 Package / container Leak orange juice There was a stain on the package that seemed to be caused by a leak. Is it safe to drink?

Source	Date	Title
Windows file system	1/3/08	milk chocolate - Ads

00000004 milk chocolate - Ads 2008-01-03 1199359192093 Ads TV ads milk chocolate I love the ads for the new milk chocolate. Could you tell me the name of the actor in the commercial?

Source	Date	Title
Windows file system	1/3/08	vanilla ice cream - Prank

00000005 vanilla ice cream - Prank 2008-01-03 1199359192312 Prank Open vanilla ice cream The cup looked like it was already opened. Is it safe?

Source	Date	Title
Windows file system	1/3/08	chocolate ice cream - Taste / smell

00000006 chocolate ice cream - Taste / smell 2008-01-03 1199359192343 Taste / smell Change chocolate ice cream It was a little more sour than when I ate it before.

Figure 6-2 Documents view showing the results based on the default search condition

6.2.1 Understanding the Documents view

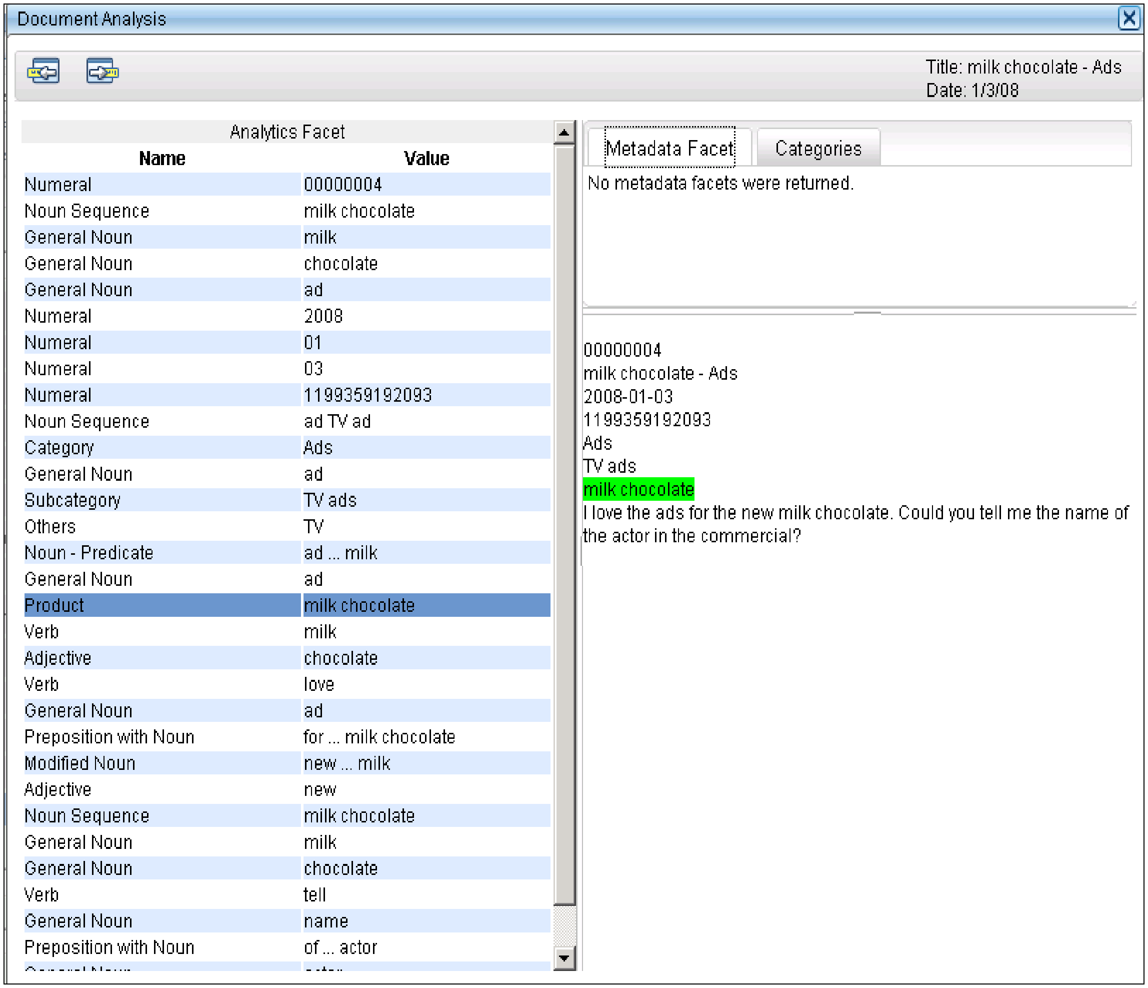
In the Documents view, you can see the following information:

- The total number of documents that match your query. In the example shown in Figure 6-2 on page 220, 852 documents are returned.
- The query that is used to produce the result. The query is hidden when the search box is hidden.
- Selected fields such as Source, Date, Title, and Thumbnails. The fields displayed as columns are based on your configuration settings on the Results Columns tab under Preferences.

- The detail of each document when you click the **Show detailed properties** icon.
- The document content when you click the link created in the Title field.

6.2.2 Viewing the document contents and facets

When you click the source icon, the Document Analysis window opens on top of the Documents view. This window contains details about your document. It shows the individual field values for the document and all annotations made to the document during text analysis. In the Document Analysis window (Figure 6-3), the Analytics Facet is listed in the left pane, and the Metadata Facet is shown in the right pane.



Title: milk chocolate - Ads
Date: 1/3/08

Analytics Facet

Name	Value
Numeral	00000004
Noun Sequence	milk chocolate
General Noun	milk
General Noun	chocolate
General Noun	ad
Numeral	2008
Numeral	01
Numeral	03
Numeral	1199359192093
Noun Sequence	ad TV ad
Category	Ads
General Noun	ad
Subcategory	TV ads
Others	TV
Noun - Predicate	ad ... milk
General Noun	ad
Product	milk chocolate
Verb	milk
Adjective	chocolate
Verb	love
General Noun	ad
Preposition with Noun	for ... milk chocolate
Modified Noun	new ... milk
Adjective	new
Noun Sequence	milk chocolate
General Noun	milk
General Noun	chocolate
Verb	tell
General Noun	name
Preposition with Noun	of ... actor
General Noun	actor

No metadata facets were returned.

00000004
milk chocolate - Ads
2008-01-03
1199359192093
Ads
TV ads
milk chocolate
I love the ads for the new milk chocolate. Could you tell me the name of the actor in the commercial?

Figure 6-3 Document Analysis window

When you select the Analytics Facet, the corresponding keywords are highlighted in green in the structured and unstructured fields of your document where the values occur. For example, we select the Product analytics facet called "`milk chocolate`", as shown in Figure 6-3 on page 221. As a result, the keyword "`milk chocolate`" is highlighted in green within the Metadata Facet pane. This function helps you to understand the data that determined the facet.

6.2.3 When to use the Documents view

The Documents view is useful when you want to see the details of the individual document after investigation in other views, such as the Trends view, or after searching the collection. By using the Preferences configuration, you can control how the search result is displayed in the Documents view.

6.3 Facets view

The Facets view shows a list of keywords that are displayed for a selected facet. The frequency count and correlation value accompany each keyword. This view is useful for seeing the keywords that make up a given facet in your data.

In the Facets view, frequency and correlation are shown as a bar chart. Each column is described as follows:

Keyword The entity that is associated with the selected facet. It can be a word, a pattern of text, or a fielded value.

Frequency Indicates the number of documents found that contain the given keyword.

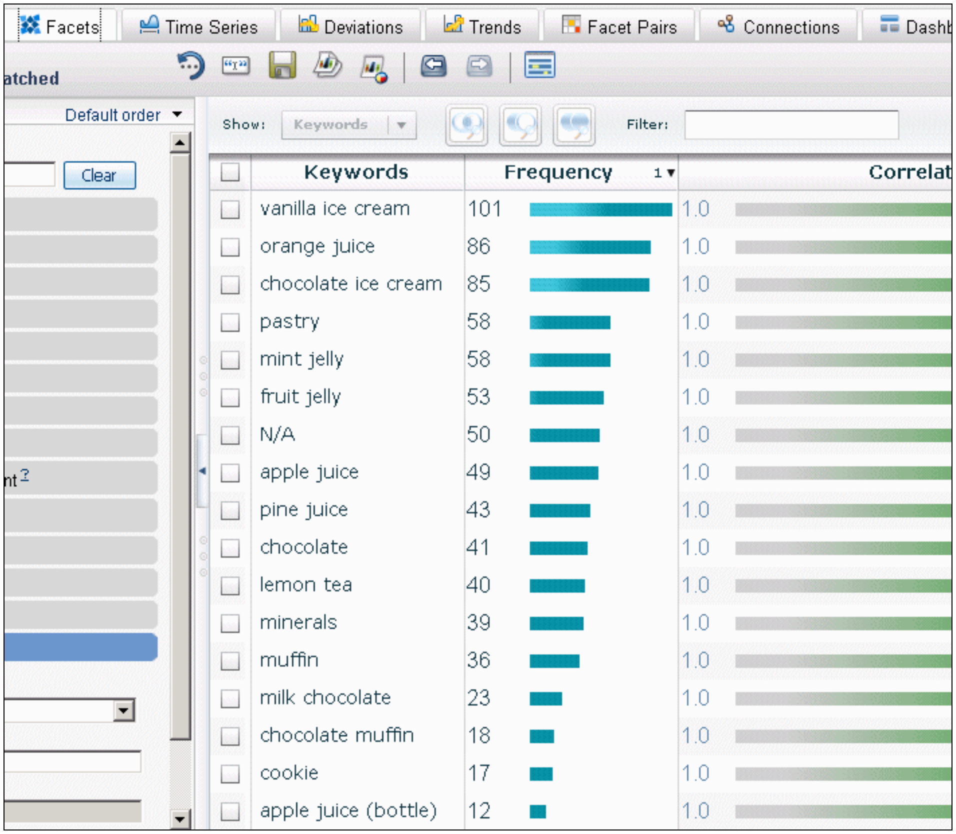
Correlation Indicates how the keyword is interrelated to the documents that are matched by your query.

You can sort the output by frequency or correlation.

Default sort order in the Facets view: By default, the documents in the Facets view are sorted by frequency in descending order. You can modify the default value by using the Preferences window.

Important: Even if you select correlation as the default sort order, the list of facets that are displayed are chosen by frequency. Therefore, if you leave the default of 100 keywords, you get only the 100 most frequent keywords. This means that sorting by correlation will *not* include values that are more highly correlated but not in the top 100 by frequency. This method of sorting is by design. The rationale is that it, if the frequency of the keyword is low, the results is not of interest.

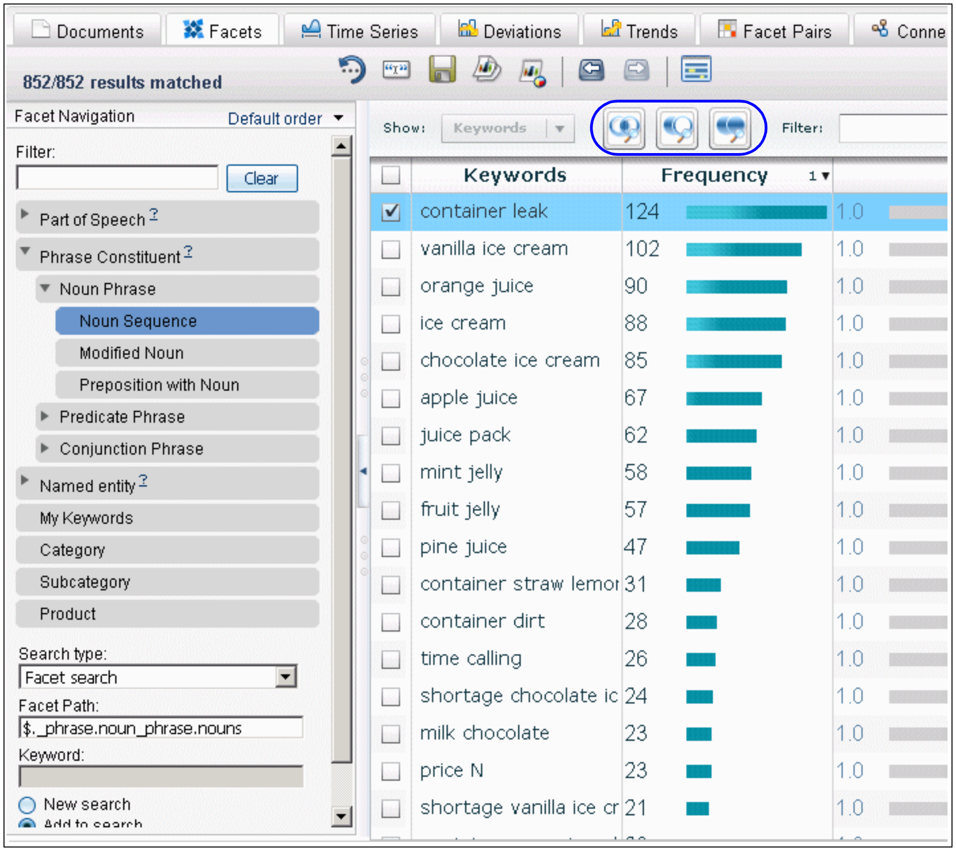
Figure 6-4 shows the view when you select the Product facet from the Facet Navigation pane. You can select any facet that you configured.



□	Keywords	Frequency 1 ▼	Correlat
□	vanilla ice cream	101	1.0
□	orange juice	86	1.0
□	chocolate ice cream	85	1.0
□	pastry	58	1.0
□	mint jelly	58	1.0
□	fruit jelly	53	1.0
□	N/A	50	1.0
□	apple juice	49	1.0
□	pine juice	43	1.0
□	chocolate	41	1.0
□	lemon tea	40	1.0
□	minerals	39	1.0
□	muffin	36	1.0
□	milk chocolate	23	1.0
□	chocolate muffin	18	1.0
□	cookie	17	1.0
□	apple juice (bottle)	12	1.0

Figure 6-4 Facets view with the Product facet selected

You can limit the scope of your analysis to one or more keywords by selecting them using their corresponding check boxes and adding them to the query. When you select a keyword, the Boolean search operator (AND, OR, and AND NOT) icons are highlighted and become active, as shown in Figure 6-5.



852/852 results matched

Facet Navigation   Default order ▾

Filter: [ ] Clear

- Part of Speech ?
- Phrase Constituent ?
  - Noun Phrase
    - Noun Sequence
    - Modified Noun
    - Preposition with Noun
  - Predicate Phrase
  - Conjunction Phrase
- Named entity ?
- My Keywords
- Category
- Subcategory
- Product

Search type: Facet search

Facet Path: $._phrase.noun_phrase.nouns

Keyword:

○ New search
● Add to search

Show: Keywords ▾   Filter:

□	Keywords	Frequency 1 ▾	
☑	container leak	124	1.0
□	vanilla ice cream	102	1.0
□	orange juice	90	1.0
□	ice cream	88	1.0
□	chocolate ice cream	85	1.0
□	apple juice	67	1.0
□	juice pack	62	1.0
□	mint jelly	58	1.0
□	fruit jelly	57	1.0
□	pine juice	47	1.0
□	container straw lemor	31	1.0
□	container dirt	28	1.0
□	time calling	26	1.0
□	shortage chocolate ic	24	1.0
□	milk chocolate	23	1.0
□	price N	23	1.0
□	shortage vanilla ice cr	21	1.0

Figure 6-5 Facets view with the search operators highlighted when a keyword is selected

Search operators: See 5.2.1, “Limiting the scope of your analysis using facets” on page 153, for details about the search operators.

After you click a specific Boolean search operator, the view is updated with the new search results and the result counts. At any time, you can go back to the Documents view to see the content of the documents that match the given query.

6.3.1 Understanding the Facets view

To effectively use the Facets view, you must understand the difference between *frequency* and *correlation*, which are described in 1.3, “Important concepts and terminology” on page 9. As a review, the *frequency value* counts the number of occurrences in documents. The *correlation value* measures the amount of uniqueness of the high frequency as compared to other documents that match your query.

Although the frequency value is useful, it might not always be as revealing as the correlation value. For example, high frequency counts for a particular model of car can be attributed to the overall popularity of the car: More cars of that model are sold than any other models. A high correlation value means that something is unusual, and further investigation and analysis are required.

When your query is set to the initial wildcard expression `*:*`, all correlation values are set to 1.0. The reason is because all of the documents in the collection are being compared to themselves, resulting in a correlation value of 1.0. After you add a keyword to your query or enter a search expression, the correlation values are recalculated and reflect the degree of uniqueness of this keyword to the other documents in the corpus that match the query.

6.3.2 Using the Facets view

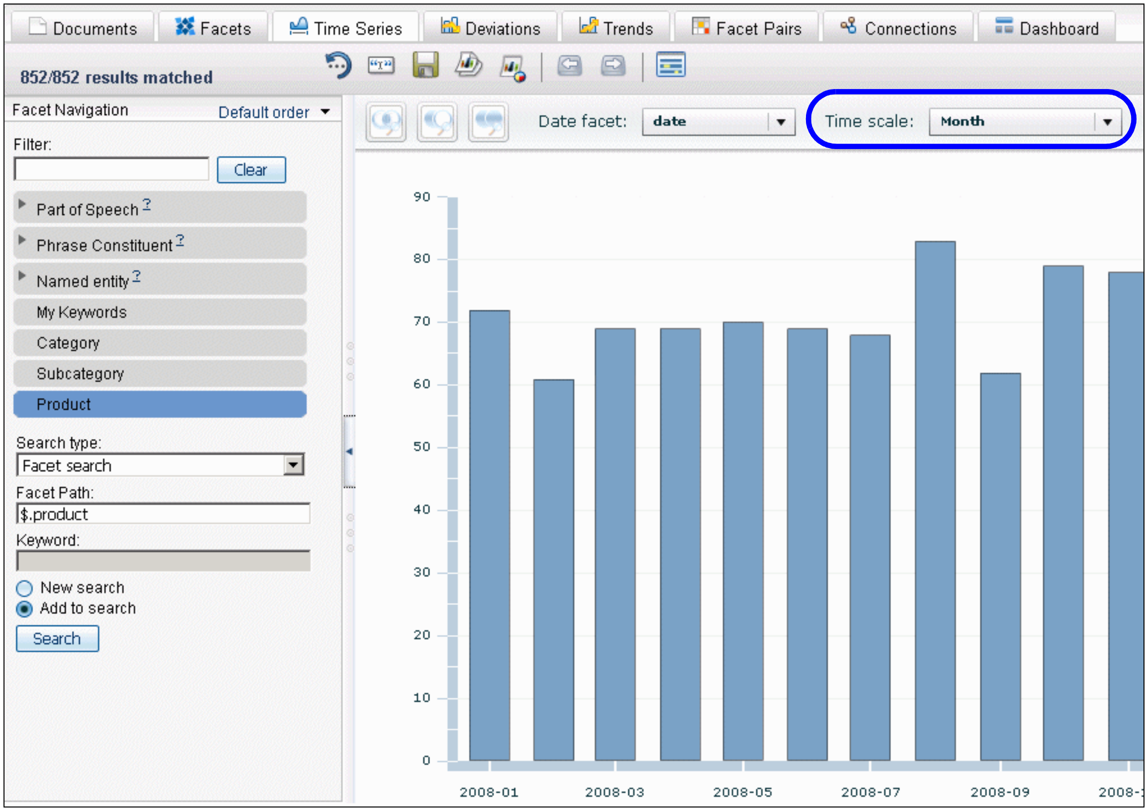
You use the Facets view when you want to see a list of keywords that are associated with a given facet. By default, the Part of Speech (POS) facet and Phrase Constituent facet are defined and populated automatically by IBM Content Analytics. Consequently, you can see various verbs, nouns, adjectives, adverbs, and phrases that are used in your text. You can use these words as keywords for dictionary and pattern rules. You can also use these words as facets to help you drill down further on a set of documents for analysis.

6.4 Time Series view

The Time Series view shows the frequency of change over time. Correlation and deviation values are not presented in this view. This view is primarily used to analyze frequency and to select a range of documents for analysis for a given time period.

The Time Series view always shows the frequency of distribution for the documents that match your query for a given period of time. For example, consider a situation where you select **vanilla ice cream** for the Product facet and add the keyword with the AND operator. In this case, you can see how the vanilla ice cream documents are distributed across the selected time scale (year, month, day) along with their computed frequencies.

In another example, the Time Series view in Figure 6-6 shows the frequency of distribution when you select the Product facet and select Month for Time scale.



852/852 results matched

Date facet: date

Time scale: Month

Search type: Facet search

Facet Path: $.product

Keyword:

○ New search
● Add to search

Figure 6-6 Time Series view showing results sorted by month

6.4.1 Features in the Time Series view

With the flexibility of the Time Series view, you can change the time scale (year, month, and day), use the Zoom in and Zoom out features, and focus on specific date ranges by using the Date facet. This section provides details about each the features of this view.

Changing the time scale

You can change the time scale of the graph when you analyze the data. From the drop-down menu, you can select the year, month, day, month of year, day of month, day of week, month of year, day of month, and day of week.

Each time scale calculates the sum of all documents for that particular time unit. For example, if you select month as the time scale, you see a bar for each month in the range of months as bounded by the documents that match your query. If your search result set spans two years, 24 months are shown.

If you select month of year, you see 12 bars; if you select day of month, you see 31 bars; and if you select day of week, you see 7 bars. For each result, you see the sum of all documents that fall on that particular date increment. For example, if you select the day of week time scale, you see seven bars, with the first bar representing the total of all documents in the result set that fall on Sunday. This feature conveniently shows the days of the week, month, or months of the year in which the most documents (or events) occur.

Zooming in and out

You can use the Zoom in and Zoom out feature especially when the bar chart becomes too busy to distinguish the exact values. As shown in Figure 6-7, you can select an area that you want to look into by dragging and zooming in on that area. After you select the area, click the **Zoom in** icon.

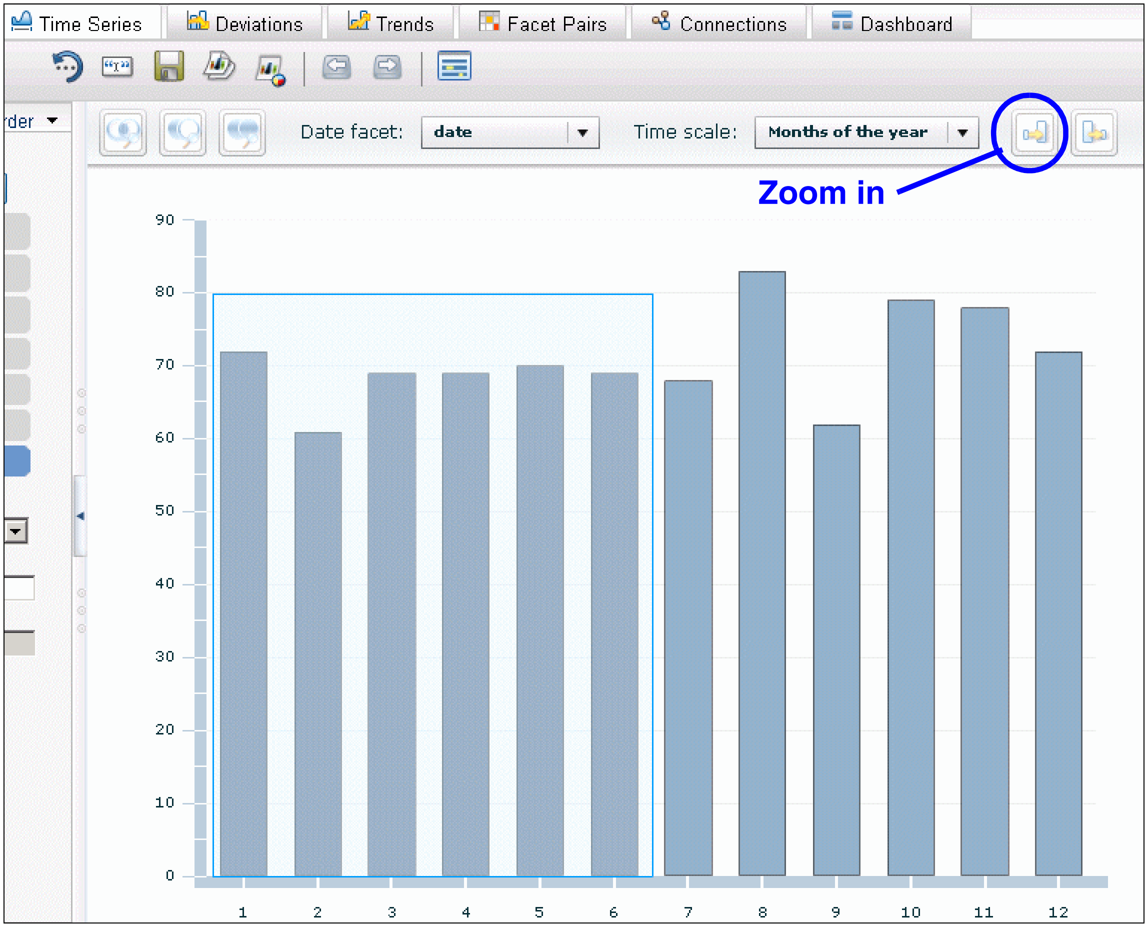


Figure 6-7 Time Series view showing the selected area to zoom in

Figure 6-8 shows the result of zooming in. When you click the **Zoom out** icon (highlighted in Figure 6-8), the view goes back to the original graph (Figure 6-7 on page 228).

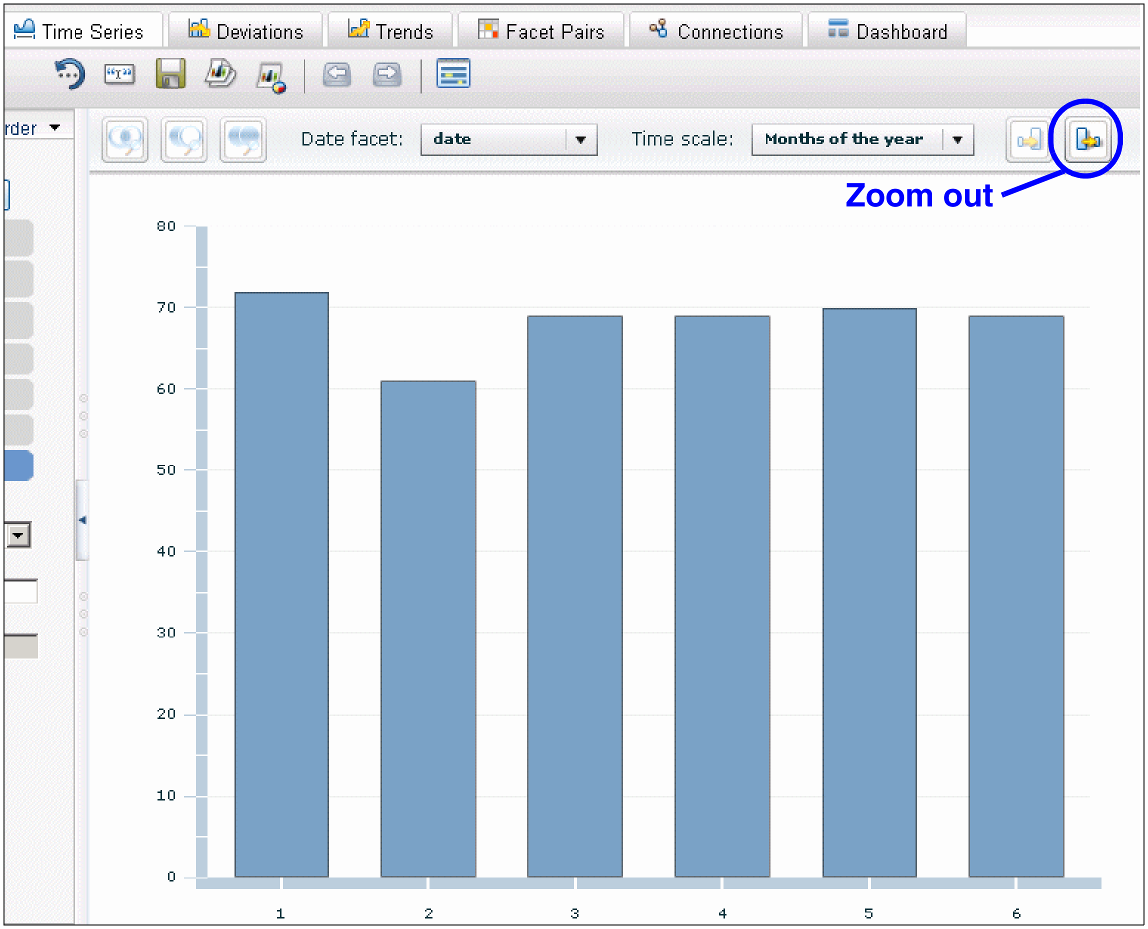
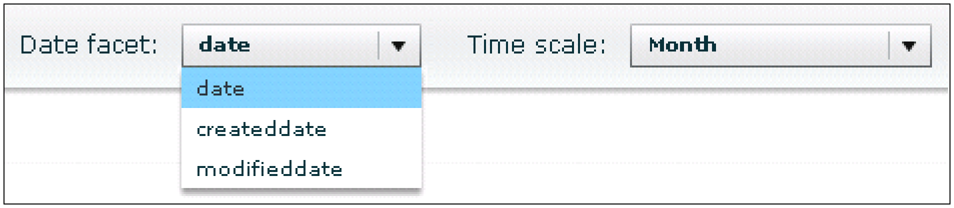


Figure 6-8 Time Series view showing the results of zooming in on the selected area

Changing the Date facet

When you configure the Date facet to contain multiple date fields, you can select a different field for the Date facet to analyze the data in the Time Series, Trends, or Deviations views. By changing the field for the Date facet, the time scale for the graph is automatically updated to use the new date field.

For example, after the Time Series view is displayed based on the reported date of the document, you might want to see the same data in the Time Series view based on the date the incident occurred to give you another analysis perspective. Figure 6-9 shows selecting the Date facet value in the Time Series view.


Date facet: date

Time scale: Month

Figure 6-9 Selecting the Date facet in the Time Series view

Configuring multiple date facets: See "Optional: Configuring the date facet" on page 113 to configure the date facet.

6.4.2 Understanding the Time Series view

The Time Series view shows the distribution of documents that match your query over a period of time. The y-axis shows the number of documents (frequency), and the x-axis shows the time scale that you selected.

When you hover the mouse cursor over a particular bar in the chart, a pop-up window opens that details the data for that unit, namely the frequency and specific date value.

6.4.3 Using the Time Series view

The Time Series view helps you to see how frequent your documents change over time and on specific months or days. The distribution always represents the documents that match your current query. By building separate queries for different combinations of facets (or search expressions), you can start to compare their frequencies of occurrences. This view also helps you to identify a frequency trend such as sudden increases for the selected facet.

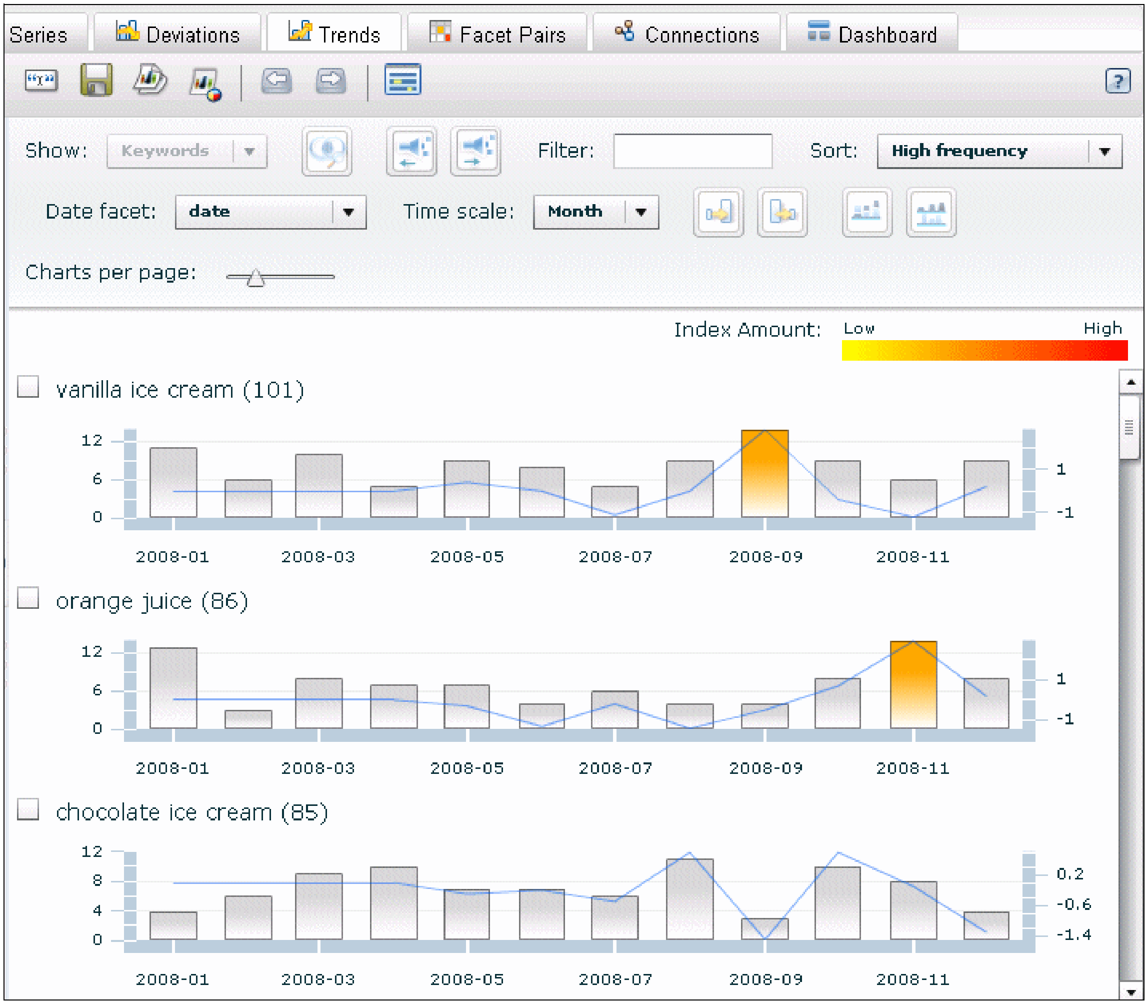
Frequency counts alone, while useful, might not be that revealing as discussed earlier. The Trends and Deviations views are better suited for this task, and they take into account time and are much more useful in spotting anomalies in your data. The Time Series view is more useful for selecting a specific date or a range of dates to limit the scope of your analysis.

6.5 Trends view

The Trends view shows sharp and unexpected increases in frequency of a facet over time. The Trends view is similar to the Time Series view in that it shows the frequency distribution of documents as a bar graph across a given time frame. You can change the time scale to year, month, or day. It also provides the same zoom in and zoom out functionality as the Time Series view. However, the Trends view has significant differences in helping you to gain additional insight from your data.

First, the Trends view requires the selection of a facet from the Facet Navigation pane on the left side. You do not need to add the facet to your query. After you select a facet, the Trends view shows a list of individual bar graphs, one for each keyword of the selected facet. Each individual bar graph behaves similarly to the Time Series graph, but also highlights trends in your data that deviate from the normal distribution.

You can use this view to analyze future predictions of sharp increases in frequency. Figure 6-10 shows the Trends view when you select the Product facet and default date facet with the month time scale.



Show: Keywords

Filter:

Sort: High frequency

Date facet: date

Time scale: Month

Charts per page:

Figure 6-10 Trends view with the Product facet, sorted by high frequency and month

6.5.1 Features in the Trends view

The following features are the same as those features described in “Features in the Time Series view” on page 227:

- Changing the time scale
- Zooming in and out
- Changing the Date facet

The time scale options do not include the ability to select the month of the year, day of the month, or day of the week. The Trends view also includes the features in the following sections.

Changing the Charts per page indicator

You can change the number of charts per page by sliding the bar as highlighted in Figure 6-11. This feature is helpful when you want to view and compare multiple charts at a time on a page or focus on only one chart. The size of the chart varies depending on the number of charts viewed per page.

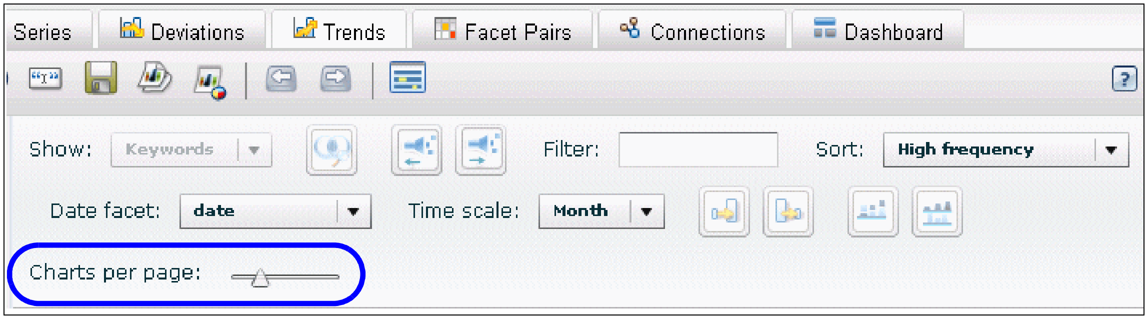
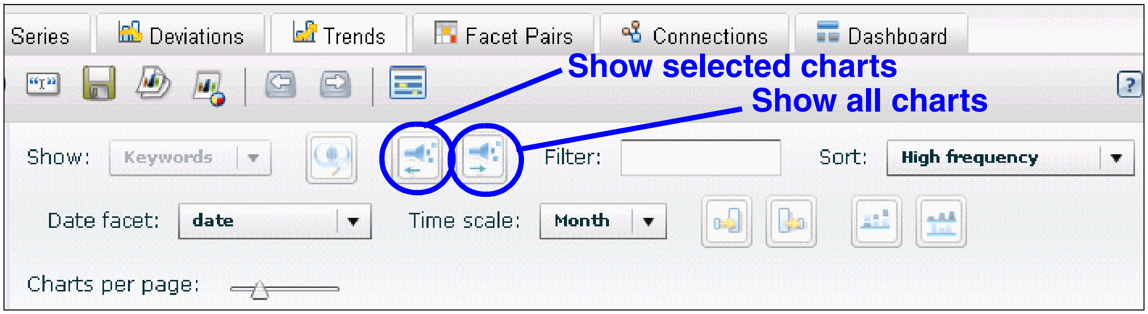


Figure 6-11 Trends view showing the Charts per page indicator

Showing selected charts or showing all charts

Each individual bar chart comes with its own selection check box. The **Show selected charts** icon (highlighted in Figure 6-12 on page 234), when clicked, reduces the view of charts to only those charts that are selected. For example, you might have a total of 48 charts, and eight charts are shown per page on six pages. You are only interested in seven charts scattered across various pages. By selecting the charts you are interested in using the check box and clicking the **Show selected charts** icon, only the six selected charts are shown on a single page for comparison.

You can revert to the original chart view by clicking the **Show all charts** icon (also highlighted in Figure 6-12).



Show selected charts

Show all charts

Show: Keywords

Filter:

Sort: High frequency

Date facet: date

Time scale: Month

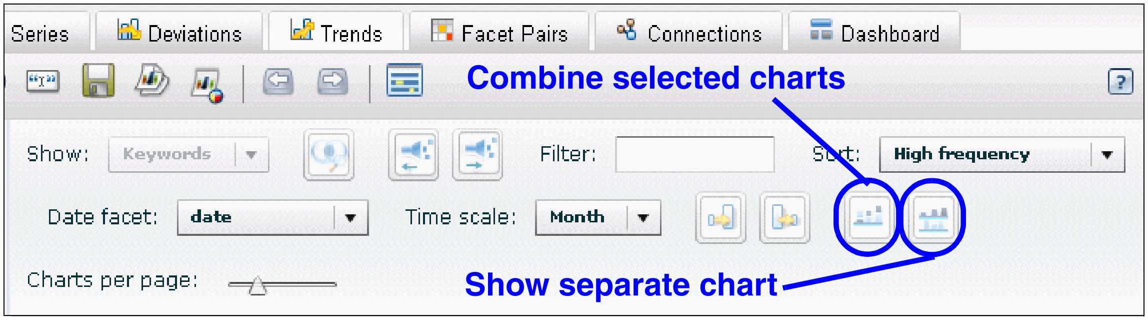
Charts per page:

Figure 6-12 Trends view showing the Show selected charts and Show all charts icons

Combining selected charts or showing separate charts

You can combine multiple selected charts into one chart by clicking the **Combine selected charts** icon (highlighted in Figure 6-13). This function aggregates all the charts into a single chart with each keyword given a different color.

You can revert to the original chart view by clicking the **Show separate charts** icon (highlighted in Figure 6-13).



Combine selected charts

Show: Keywords

Filter:

Sort: High frequency

Date facet: date

Time scale: Month

Charts per page:

Show separate chart

Figure 6-13 Trends view showing the Combine selected charts and Show separate charts icons

Filtering the result by keyword

When you type a keyword in the Filter field box, the charts whose keywords contain the input filter are displayed in the view. The view is updated dynamically, so that you can see the filter result immediately.

6.5.2 Sort criteria

You can select from the following sort criteria to assist in the investigation of your data:

- Highest frequency (default)

  This criterion lists, in descending order, those graphs with the highest frequency counts. The keywords graph at the top of the list contains a time period with the greatest frequency count. This method offers a quick way to see which keywords contain the most occurrences. Remember, frequency counts are not always as informative, but rather expected. For example, when looking at the sales of snow shovels, you expect a high number to be sold during the winter months.

- Highest index

  This criterion lists, in descending order, those graphs with the highest deviations from their expected averages within the current time frame selected. This method offers a quick way to see which keywords are operating out of the norm the most.

  For example, you might have a facet that tracks the frequency of car part failures. The highest index might list, at the top of the list in descending order, the car part with the highest deviation (index) from its expected trending average from any other car part (and their deviations). The highest index can occur anywhere in the time period selected as opposed to the last index, which only focuses on the last time increment in the series.

- Latest index

  This criterion is similar to the highest index criterion but only uses the latest time unit (year, month, or day) to determine which keywords have the highest index. This method is a quick way to see which keywords most recently are operating out of the norm.

- Name

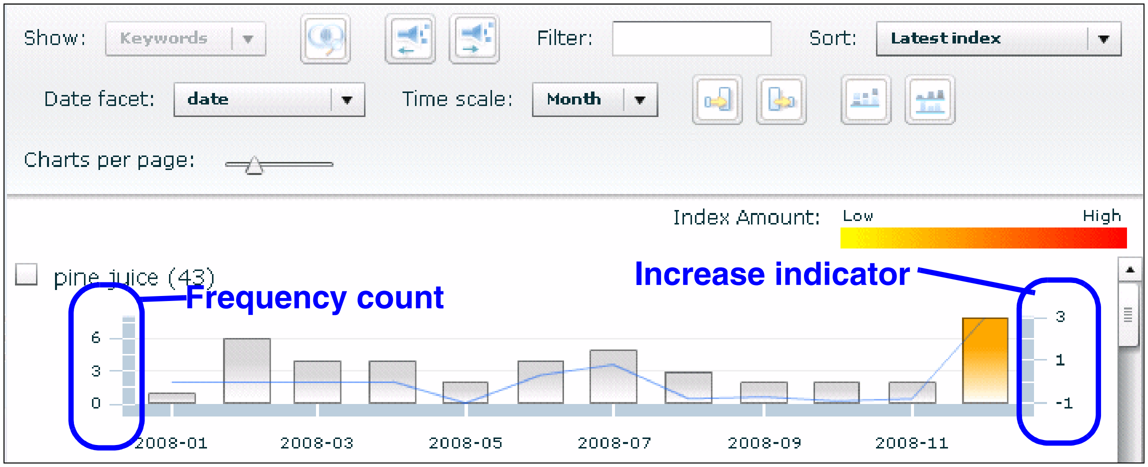
  This criterion alphabetically sorts the graphs by name in either ascending or descending order. This method is a quick way to find particular keywords that you are interested in.

6.5.3 Understanding the Trends view

In the Trends view (Figure 6-14), the frequency of the selected time period is shown as a bar chart. It is scaled accordingly from 0 to the maximum frequency count along the vertical y-axis on the left side.

On the right side of the y-axis is a scale that represents the increase indicator. The *increase indicator* is a scale to measure the increase ratio of the frequency for a given time interval as compared to the expected average frequency that is calculated based on the changes in the past time interval frequencies. This expected change in frequency is estimated by using a modified Poisson distribution. The increase indicator is shown as a blue line graph in the chart, as shown in Figure 6-14.

The bar chart is in color (to highlight) whether the frequency deviation is higher than what was expected within reasonable limits. This result means that the actual value is greater than the estimated value by a certain amount. Figure 6-14 shows a brighter orange color for December 2008. As you can see from the blue line in the graph, the increase indicator shows a sudden jump, which means that you might want to conduct additional investigation and analysis for that time period.


Show: Keywords

Filter:

Sort: Latest index

Date facet: date

Time scale: Month

Charts per page:

Figure 6-14 Trends view: Pine juice with frequency count and increase indicator

Calculating the index in the view

How is the increase indicator that is shown as a blue line in the chart calculated? To calculate the index (that is, the increase indicator), you must calculate the average global frequency, which is the average frequency of all searched documents over the given time period. You must also calculate the average keyword frequency over the same period.

In simple terms, to calculate an average, you add all the frequencies together and divide by the number of dates in the series. Because this calculation is too simplistic, you also want to account for the *decay factor*. *Decay* means that you want the past to become decreasingly relevant, the more distant it gets. That is, each frequency that is calculated is weighted according to a decay constant. It is a matter of both adding all the frequencies together and equally contributing values in the average calculation. The decay factor is applied to each frequency first. Thus, if the decay value is 0.85 (the default setting for the increase indicator in the text miner application), the frequency of the *n*th–4th date contributes less (about half) to the calculation as the frequency for the *n*th date. Both the global and the keyword time series average frequencies are weighted this way.

With the weighted average frequency time series, you can accurately estimate the frequency count for a future date. That is assuming that the time series is constant and you factor in the variation within all searched documents using the weighted global average frequency. This estimate gives you the ability to collate the increase index, to the degree to which the frequency of this keyword has increased or will increase, for a particular date.

Given this description of how the expected frequency index is calculated, it is obvious why the first four time intervals of all your graphs show no change and are flat with no highlighted colors showing. The algorithm needs the first four intervals to start the calculation of the Poisson distribution.

6.5.4 When to use the Trends view

With the Trends view, you can detect sharp and unexpected increases in the frequency of a given keyword. Usually a sharp increase indicates that you need to do more investigation. If the increase indicator index is higher than the threshold, the bar chart is in color, so that it is easier for you to discover the anomaly.

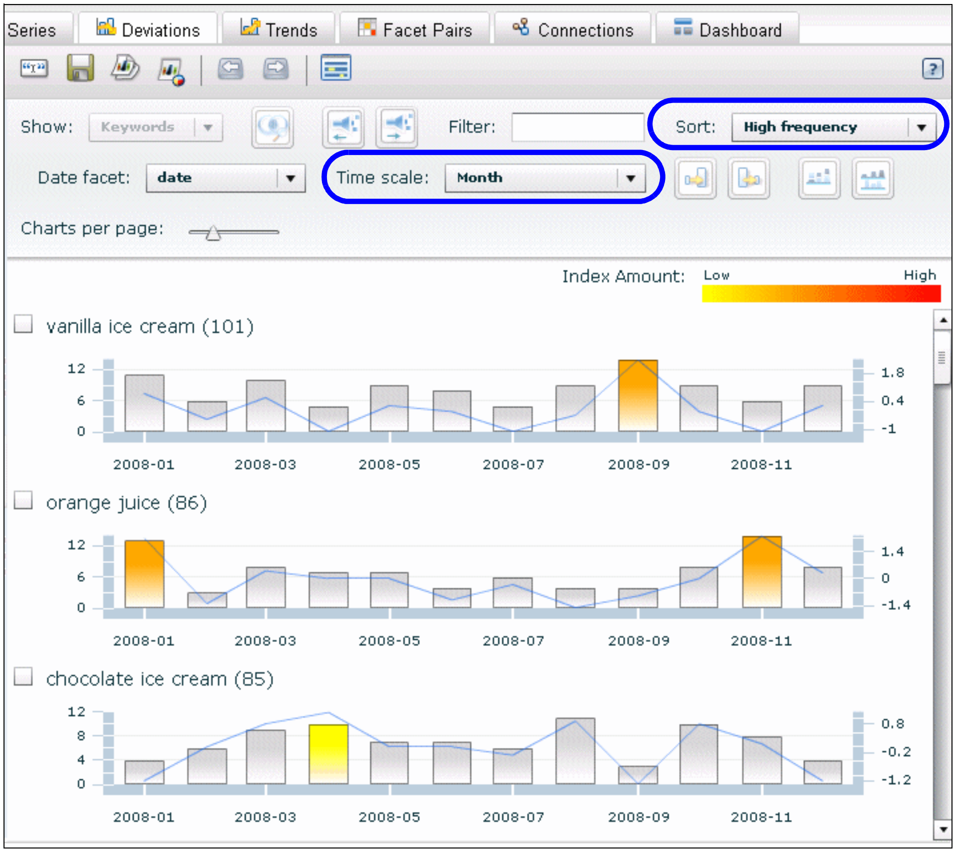
6.6 Deviations view

The Deviations view shows the deviation of keywords for a given time period. The Deviations view is similar to the Trends view in that it requires the selection of a facet and shows the corresponding individual graphs for each keyword. The controls across the top function are the same as in the Trends view with one exception. Three additional selections are available from the time scale pull-down menu, namely the month of year, day of month, and day of week. These additional selections provide greater insight into cyclic changes in your data.

The greatest difference between the two views is what the graphs are trying to convey when certain bars are highlighted, indicating something of interest. In particular, the Trends view alerts you when a keyword is trending up or down by an unexpected amount, and the expected amount is calculated based on the past history of frequency changes. This view is more focused on the trending of frequency counts over time.

The Deviations view is focused on how much the frequency of a given keyword deviates from the expected average for the given time period (not previous periods). The expected average takes into account all the averages of the other frequency counts for the given time period. The Deviations view is useful for identifying patterns that occur cyclically and alerts you when those cyclic patterns have an unexpected change. You can use this view to show seasonal patterns in your data or patterns that occur on a monthly or weekly basis.

Figure 6-15 shows the Deviations view when you select the Product facet with Time scale set to Month, the Date facet set to the default of date, and Sort set to High frequency.



Show: Keywords
Filter:
Sort: High frequency
Date facet: date
Time scale: Month
Charts per page:

Figure 6-15 Deviations view showing sorting by high frequency and month

6.6.1 Features in the Deviations view

The following features are functionally the same as the features for the Trends view:

- Changing the Charts per page indicator
- Showing selected charts or showing all charts
- Combining selected charts or showing separate charts
- Zooming in and out

- Changing the Date facet
- Sort criteria
  - High frequency
  - High index
  - Latest index
  - Name (ascending)
  - Name (descending)

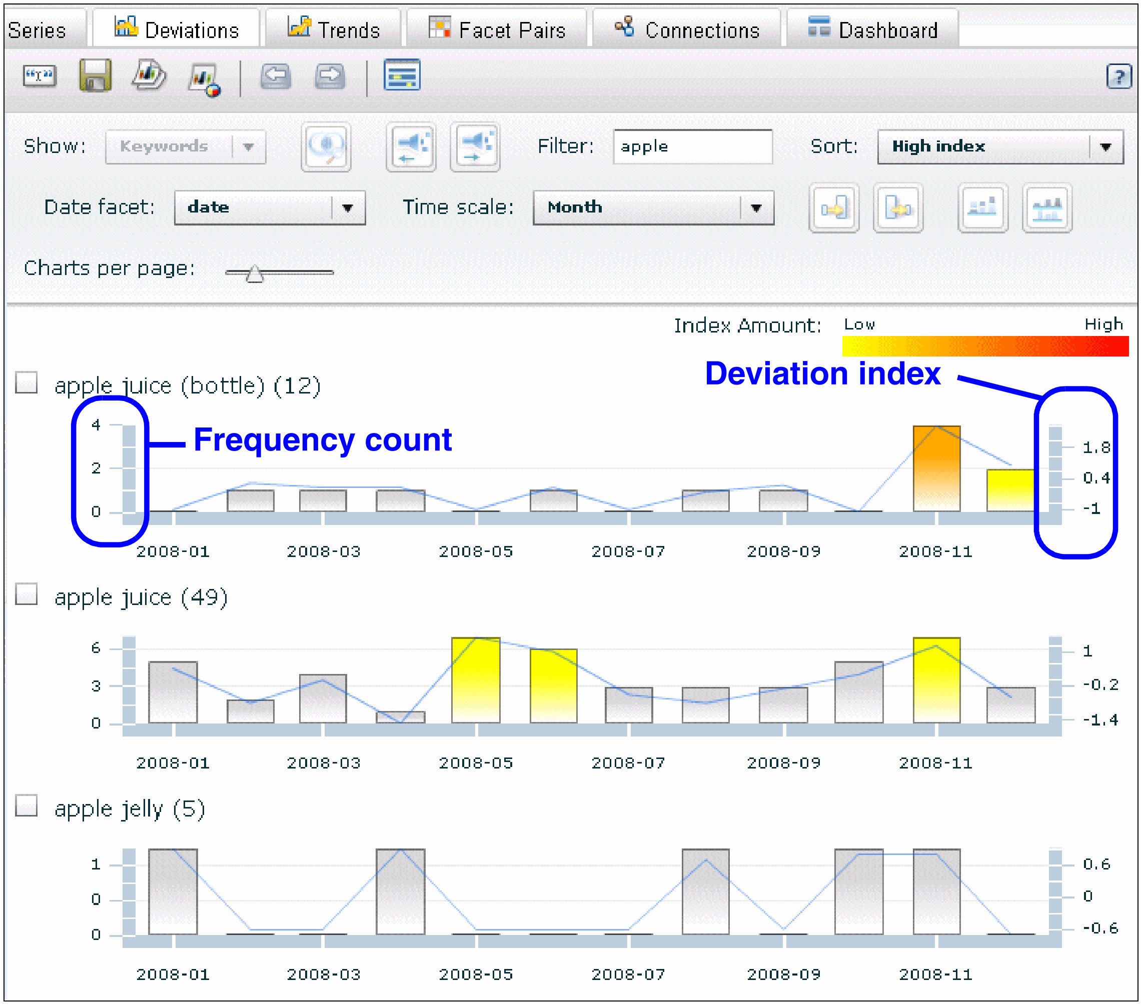
See 6.5, “Trends view” on page 231, for a detailed description of their functionality. Notice that the time scale feature in the Deviations view is identical to the same function in the Time Series view. As in the Time Series view, this time scale includes options for selecting the month of year, day of month, and day of week. You select these options when you want to see seasonal changes or monthly and weekly changes.

6.6.2 Understanding the Deviations view

In the Deviations view (Figure 6-16 on page 241), the frequency of the selected time period is shown as a bar chart and measured at the y-axis on the left side. The deviation index score is measured at the y-axis on the right side. The *deviation index score* is the standardized residual. It is referred to as *index* in the chart.

The deviation index score indicates how the actual value deviates from the expected value for a given time frame. The blue line in the chart shows the deviation index scores. The bar chart is displayed in color if the deviation index score is higher than the threshold, which indicates that the actual value is greater than the expected value.

For example, we select the Product facet and filter with the keyword “apple,” choose **High index** for Sort, and select **Month** for Time scale, as shown in Figure 6-16.


Series | Deviations | Trends | Facet Pairs | Connections | Dashboard

Show: Keywords

Filter: apple

Sort: High index

Date facet: date

Time scale: Month

Charts per page:

Figure 6-16 Deviations view showing the frequency count and deviation index

In the graph, the bar chart in 2008-11 (November 2008) for apple juice (bottle) is highlighted with orange, which indicates that the index amount is relatively high. Also, notice the yellow highlighted bars for 2008-12 for apple juice (bottle), and the graph below apple juice for the months 2008-05, 2008-06, and 2008-11.

Also notice that the frequency for apple juice (bottle) on 2008-11 is highlighted as orange with a frequency of 4. This frequency is less than the frequency of 7 for apple juice on the same month 2008-11, which is highlighted in yellow. This result occurs because the deviation index score for apple juice (bottle) is higher than for apple juice and warrants the stronger red color.

How is the deviation index score calculated? It is calculated by using the frequency counts of each keyword in the given time period and the total number of frequency counts during the given time period.

If you go back to the Time Series view, select the **Product** facet, and select the Month as time scale, you see that the following results:

- The total number of documents (all keywords) for 2008-11 is 78 from the Time series view.
- The total frequency count (for the entire time period) for apple juice (bottle) is 12 from the Deviations view.
- The total frequency count (for the entire time period) for apple juice is 49 from the Deviations view.
- The frequency of 2008-11 for apple juice (bottle) is 4 from the Deviations view.
- The frequency of 2008-11 for apple juice is 7 from the Deviations view.

The expected value for apple juice (bottle) and apple juice is calculated based on those values. It indicates which value is expected for the frequency of the keyword statistically. That is, the expected value multiplies the actual frequency of the selected keyword by the ratio of the selected keyword frequency in the entire frequency.

In this case, the expected value of 2008-11 for apple juice (bottle) is 1.0. (For 12 months, the total frequency count is 12, and therefore, the monthly count is about 1.0.) The expected value of 2008-11 for apple juice is 4.5. (For 12 months, the total frequency count is 49. Therefore, the monthly count is a little more than 4.)

The actual frequency value of 2008-11 for apple juice (bottle) is 4, which is greater than the expected value of the keyword “apple juice (bottle),” which is 1.0. Its delta ratio is greater than the one for apple juice. Notice that we do not compare the delta of the actual frequency and the expected value. Content Analytics calculates the deviation index score itself based on these values.

The bar chart is in color based on the deviation score index so that you can easily determine which keyword in the selected facet is worth further investigation.

6.6.3 Using the Deviations view

The Deviations view is helpful when you want to see the deviation of the keyword for the selected facet within the given time period, such as month and day of the month. For example, you might want to see if the characteristics between Monday and Wednesday have any noticeable change when you look at the Product facet with day of the week selected as the time scale.

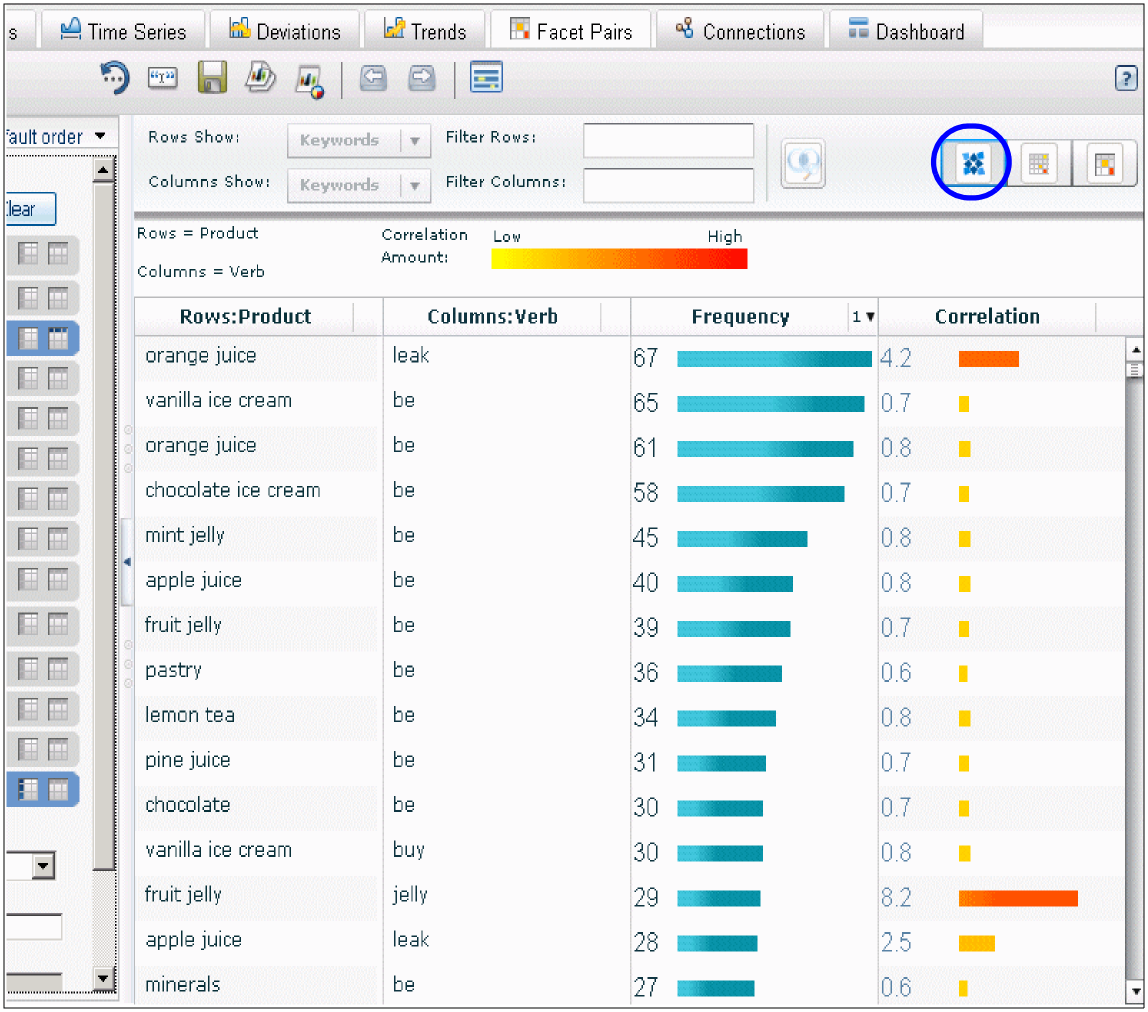
You can compare the deviation within the selected facet (that is the selected aspect). You can also see if you can find anything noticeable in that aspect with the given time scale, compared to the keywords that are found within the facet. In an earlier example, the deviation score index of 2008-11 for apple juice (bottle) is greater than the one for apple juice when you look at the data with the month time scale. Therefore, you might want to drill down the documents related to apple juice (bottle) and investigate why its deviation is noticeable compared to the other product found in the Product facet at 2008-11.

6.7 Facet Pairs view

The Facet Pairs view (Figure 6-17 on page 244) shows the correlation of keywords from two selected facets. In this view, you select two facets from the Facet Navigation pane to see the correlation of these facets.

After you select two facets to analyze, you can also choose from the following three alternative displays of the facet pair comparison:

- Table view
- Grid view
- Bird's eye view



Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns:

Rows = Product

Columns = Verb

Rows:Product	Columns:Verb	Frequency 1 ▾	Correlation
orange juice	leak	67	4.2
vanilla ice cream	be	65	0.7
orange juice	be	61	0.8
chocolate ice cream	be	58	0.7
mint jelly	be	45	0.8
apple juice	be	40	0.8
fruit jelly	be	39	0.7
pastry	be	36	0.6
lemon tea	be	34	0.8
pine juice	be	31	0.7
chocolate	be	30	0.7
vanilla ice cream	buy	30	0.8
fruit jelly	jelly	29	8.2
apple juice	leak	28	2.5
minerals	be	27	0.6

Figure 6-17 Facet Pair view showing the Verb facet and the Product facet in Table view format

6.7.1 Table view

As shown in Figure 6-17, the Table view shows the selected two facets using a table style. By default, it is sorted by frequency. In this view, it is important to focus on which pair has the highest correlation value.

In Figure 6-17, we select the **Product** facet for Rows and the **Verb** facet for Columns and sort the result by frequency, not by correlation. As a result, the frequency of the document that contains both the keyword "orange juice" and "leak" is 67, and the correlation of those selected keywords is "4.2." The Table view is most useful when you sort by frequency.

The Table view is the default view when you initially select two facets to compare. You can change the default Facet Pairs view to another view format in the Preference window.

6.7.2 Grid view

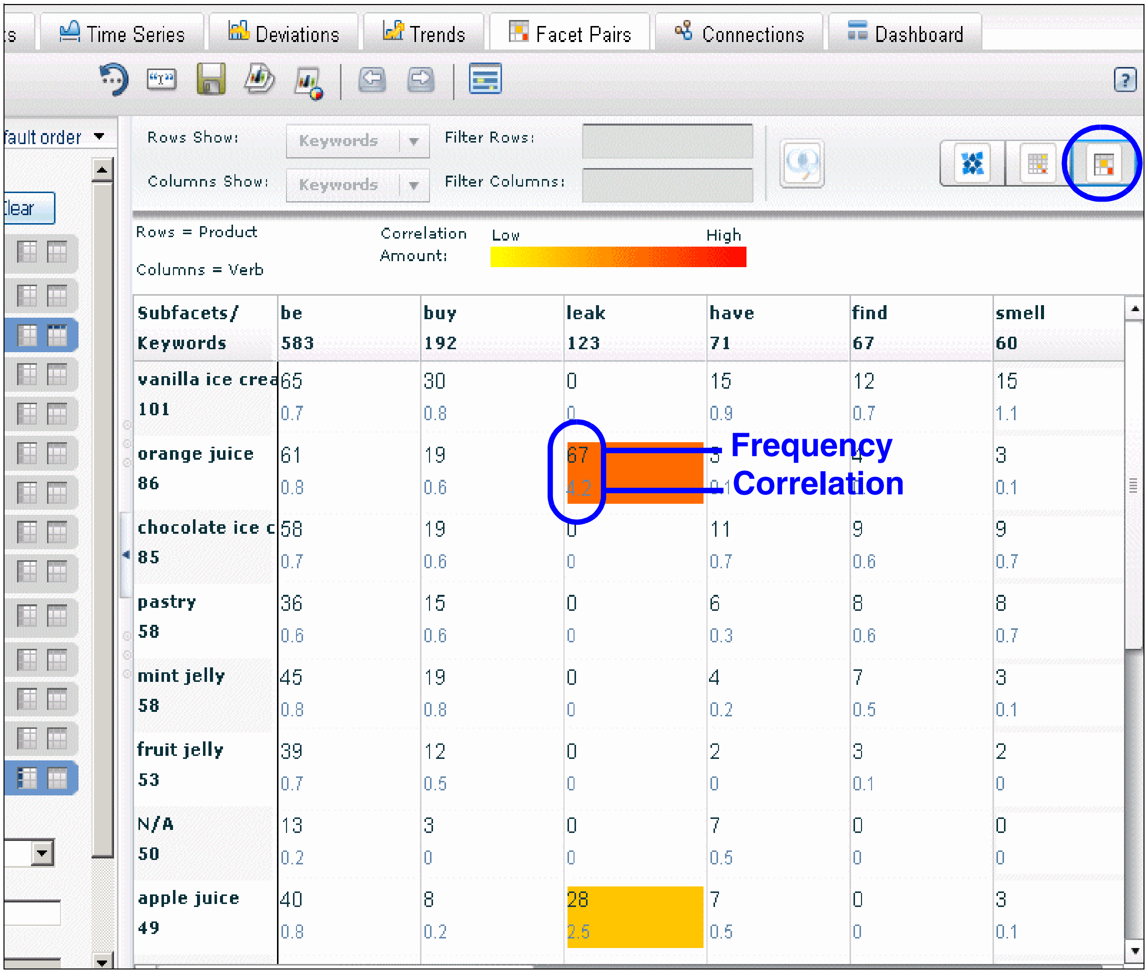
You can see the correlation for the selected facets by using the Grid view. In the example shown Figure 6-18 on page 246, we select the **Product** facet for Rows and **Verb** facet for Columns. The cell that is the intersection of orange juice and leak is highlighted in orange, which indicates that the correlation value is greater than the threshold when compared to the other correlation values. That is, the leak issue has a high correlation with the orange juice product, or the orange juice product has a high correlation with the leak issue.

You might notice that two values found in the orange cell (67 and 4.2) are the same as those shown in the Table view in Figure 6-17 on page 244. The first row in the cell is the frequency (the number of documents) that contains both the selected keywords, and the second row in the cell is the correlation value.

You also see the numbers, such as 86 under the keyword "orange juice" and 123 under the keyword "leak." These values are the frequency of each keyword and the number of documents returned by the selected keyword.

In this view, you can see the comparison values in table form by row and column, one facet for each dimension. The Grid view can calculate 100 x 100 cells of the table data, based on the highest frequency. Consequently, the top 100 most frequent keywords in one selected facet are displayed as rows. Also the top 100 most frequent occurring keywords in the other selected facet are displayed as columns.

What if either of the selected facets has more than 100 keywords? The Facet Pairs view is constructed based on the assumption that the users want to see the highest frequency first because they usually contain the most interesting correlations. If you must oversee all of the data, instead consider using deep inspection, which is explained in 10.7, “Deep inspection” on page 431. Or, you can confirm the keyword connection by using the Connections view as explained in 6.8, “Connections view” on page 250.



Rows Show: Keywords Filter Rows:

Columns Show: Keywords Filter Columns:

Rows = Product

Columns = Verb

Correlation Amount: Low High

Subfacets/ Keywords	be 583	buy 192	leak 123	have 71	find 67	smell 60
vanilla ice crea 101	65 0.7	30 0.8	0 0	15 0.9	12 0.7	15 1.1
orange juice 86	61 0.8	19 0.6	67 4.2	3 0.1	4 [illegible]	3 0.1
chocolate ice c 85	58 0.7	19 0.6	0 0	11 0.7	9 0.6	9 0.7
pastry 58	36 0.6	15 0.6	0 0	6 0.3	8 0.6	8 0.7
mint jelly 58	45 0.8	19 0.8	0 0	4 0.2	7 0.5	3 0.1
fruit jelly 53	39 0.7	12 0.5	0 0	2 0	3 0.1	2 0
N/A 50	13 0.2	3 0	0 0	7 0.5	0 0	0 0
apple juice 49	40 0.8	8 0.2	28 2.5	7 0.5	0 0	3 0.1

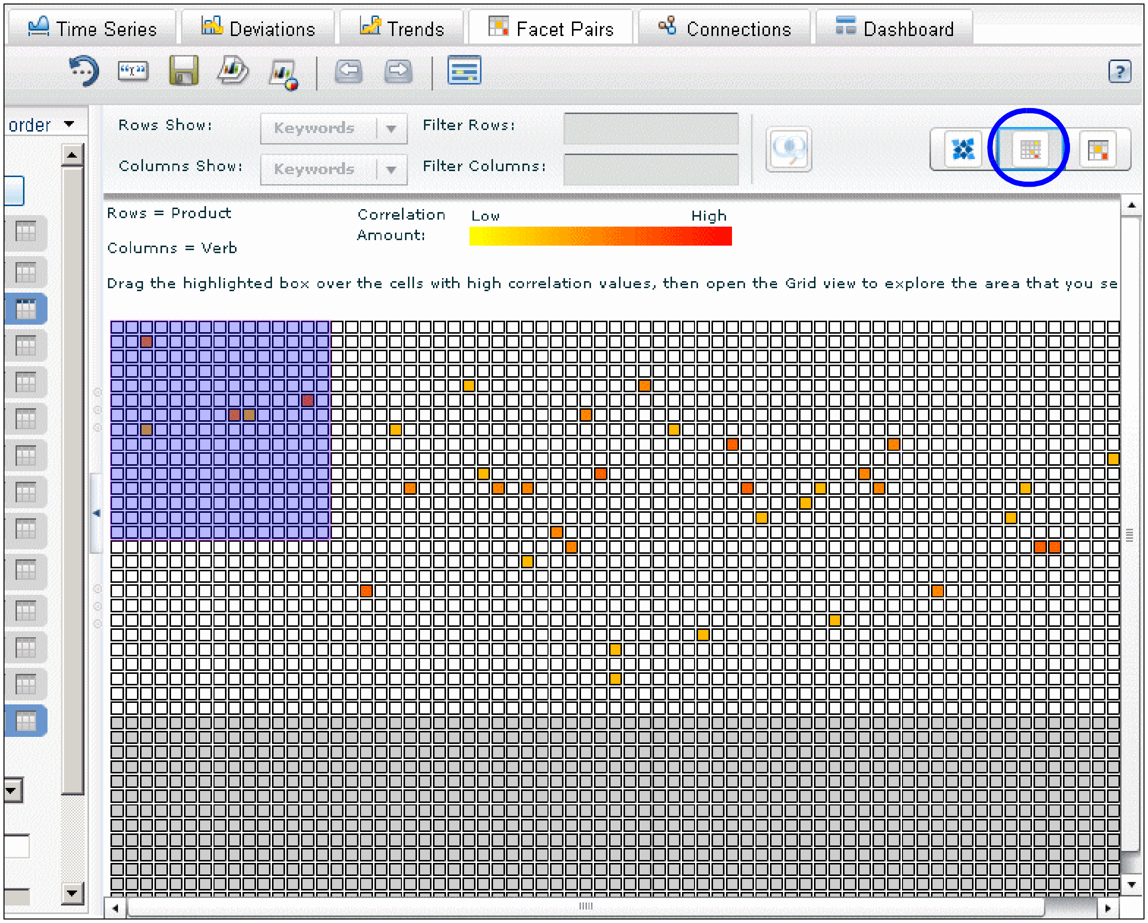
Figure 6-18 Facets Pairs view showing the Verb facet and Product facet in the Grid view format

6.7.3 Bird’s eye view

By default, in the Grid view, you can only view a 15 x 15 celled area of the 100 x 100 table at a time. The use of the 15 x 15 viewing area is implemented for performance reasons to keep the number of required calculations low. With the Bird’s eye view, you can select other areas of the table that you might want to see that are not in the default 15 x 15 viewing area.

As shown in Figure 6-19, the current 15 x 15 viewing area is displayed as a blue box in the upper left corner of the table. The dimensions of the table within the 100 x 100 limit are displayed as white cells. To select a different viewing area, move your cursor to where you want to start in the table, click and drag the blue box, and then click the table or the grid view to see the area.

Notice that the values in the individual cells (keywords, frequency, and correlation values) are displayed in a pop-up box as you hover your cursor over the cell. Also notice how the colors of highly correlated cells are maintained in the bird's eye view to provide a convenient way to quickly locate areas of interest in the 100 x 100 table.



Drag the highlighted box over the cells with high correlation values, then open the Grid view to explore the area that you se

Figure 6-19 Facets Pair view: Verb and Product facets in the bird's eye view format

6.7.4 Understanding the Facet Pairs view with correlation values

With the Facet Pairs view, you can identify a high correlation of keywords from the selected facets. Content Analytics requires two sets of search results to calculate a correlation. Accordingly, you select two facets that represent the two search result sets of the document set.

This section explains how the correlation value is calculated in Figure 6-19 on page 247 by using the Product facet as the row and Verb facet as the column.

Calculating the correlation value

To calculate the correlation value, first you must know the following information:

- The total number of occurrences of “orange juice” is 86.
- The total number of occurrences of “leak” is 123.
- The entire number of documents in the corpus is 852.

Therefore, about 14% of the documents (`123/852`) contain the keyword “leak.” This value is referred as the *density* of the keyword in a given document set.

Next, you must look at the intersection cells of the keywords, “orange juice” and “leak.” You see that the total number of occurrences of documents that includes both the keywords “orange juice” and “leak” is 67. Thus the density of the keyword “leak” in the document set for the keyword “orange juice” is 78% (that is `67/86`).

When you calculate the correlation value, you must consider the ratio of these two density values as the correlation value:

- The density of the given set of documents that includes the specific keyword
- The density of the entire set of the documents in the whole collection

That is, you are interested in the ratio of the following items:

- The density of the keyword “leak” in the document set for the keyword “orange juice” (`67/86 = 78%`)
- The density of the keyword “leak” in the entire document set (`123/852 = 14%`)

In this example, the correlation value of orange juice and leak is calculated as 5.5 (roughly `77%/14%`).

The correlation value calculated in this manner is not reliable especially when the number of documents, which includes both keywords, is relatively small. For example, the number of documents that includes keyword B is 2, while the number of documents that includes keyword C is 100. Then, consider the number of documents that include keyword A in both document sets. The density is 50% for both document sets. Which value is more reliable?

- The number of documents that includes keyword A and keyword B is 1 (50% of the document set that includes keyword B).
- The number of documents that includes keyword A and keyword C is 50 (50% of the document set that includes keyword C).

In this case, you must consider the first case (50% calculated by `1/2` is the document that includes keyword A and keyword B) is *less* reliable. Content Analytics takes into account such situations by applying a *reliability correction* that uses statistical *interval estimation* to make the calculated correlation more reliable. (Usually it makes the calculated correlation value smaller to some degree.)

With reliable correction, the earlier example correlation value of 5.4 becomes 4.2, as shown in Figure 6-17 on page 244 in the Table view. The correlation value 4.2 is higher than the normal threshold.

Interval estimation: The topic of interval estimation is outside the scope of this book.

Correlation value: In this example, the correlation value that is calculated from the given figure is much higher than the one that is displayed in the Table view. This result is normal because Content Analytics adjusts the correlation value to a smaller value to some degree by reliability correction. Remember that the data distribution is a sample data set and is not a real case. If the document set is relatively small and not reliable, the correlation value is reduced to be a more reliable value.

6.7.5 Using the Facet Pairs view

The Facet Pairs view is useful when you want to compare facets of your collection and have Content Analytics show you how highly correlated they are to each other. Content Analytics highlights intersecting cells when two keywords are highly correlated. The Bird's eye view is used to review the entire table to quickly identify these highly correlated cells. You use the Grid view to focus on that specific area of intersecting cells.

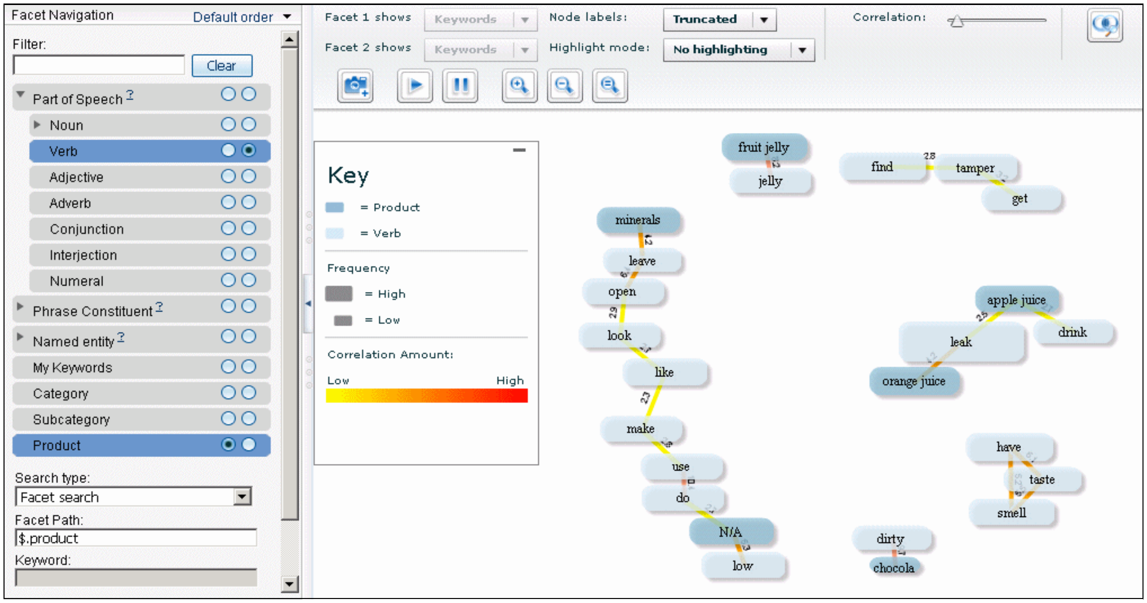
After you discover the highly correlated keyword pairs, you can go back to the Documents view and look at the textual data (content of the document) and determine if there are any possible trends or insight there.

Remember that the Facet Pairs view only concentrates on the top most frequently occurring keywords. If you need to consider all of your data, use deep inspection as explained in 10.7, "Deep inspection" on page 431.

6.8 Connections view

The Connections view shows a graphical view of the relationship between keywords or subfacets within selected facet pairs. This view is another representation of the correlation of keywords within the selected facet pairs. The keyword is represented by a node, and the link between the nodes shows the correlation value between the two keywords.

Figure 6-20 shows an example of the Connections view when you select the Product facet and the Verb facet.



Facet 1 shows: Keywords

Facet 2 shows: Keywords

Node labels: Truncated

Highlight mode: No highlighting

Search type: Facet search

Facet Path: $.product

Keyword:

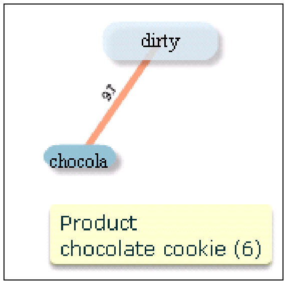
Figure 6-20 Connections view when selecting Product facet and Verb facet

> **Representation in the Connections view:** Depending on the browser window size or your operation, the representation in the Connections view can vary even though the same facet pairs are selected.

In the Connections view, you can determine the highly correlated keyword pairs by focusing on the size of the node, link color, and link length:

- The Node shows the keyword or subfacet in the selected facet pairs.
  - Node size represents the frequency of the keyword or subfacet. The larger node size represents a higher frequency count in the entire document corpus.

- Node color indicates the selected facet where the keyword is located. Our example has two node colors: light blue and dark blue. The keywords “leak”, “use,” and “find” belong to the Verb facet (light blue), while the keywords “orange juice”, “apple juice,” and “chocola” belong to the Product facet (dark blue).
- When you move the mouse pointer over a particular node, the facet name, keyword, and frequency are displayed, as shown in Figure 6-21.



Product
chocolate cookie (6)

Figure 6-21 Tooltip showing the facet name, keyword, and frequency of the node

- Link color shows the rank of the correlation index. The link in red has a higher correlation value than a yellow link color.
- Link length reflects how tightly the two nodes are correlated. The higher correlation is between two nodes, the shorter the length of the link is.

You can modify the rendering behavior of the Connections view from the **Connections** tab in the Preferences window. The following settings are among some of the settings that are configurable:

- Do not allow nodes to overlap
- Link length corresponds to correlation values
- Node size corresponds to frequency values

Configuration in Preferences: From the **Connections** tab in Preferences, you can also select the following attributes:

- Number of results to show for Facet1 per page
- Number of results to show for Facet2 per page
- Show the keywords or Sub facets by default for Facet1
- Show the keywords or Sub facets by default for Facet2

The Number of results to show for Facet1 or Facet2 is the number of facets used for the analysis. By default, 50 is set for both facets, which does not mean that 50 nodes are displayed in the search results. Fewer nodes are displayed in the search result if fewer nodes are involved in the correlation. Increasing the value can affect the performance of showing the Connections view user interface.

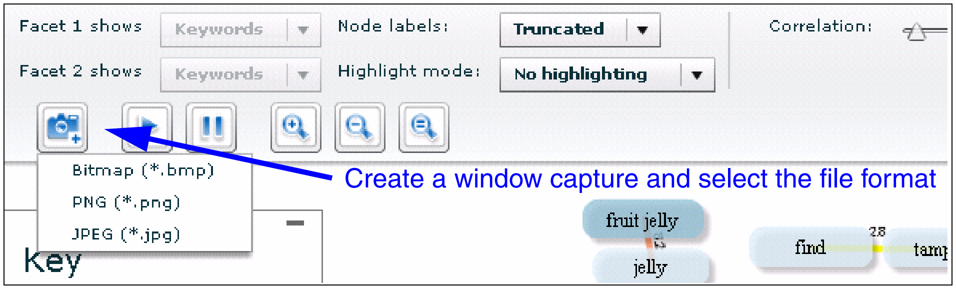
6.8.1 Features in the Connections view

The Connections view has the following features:

- Creating a window capture and saving it as an image file
- Resuming or pausing rendering
- Zoom in and Zoom out
- Zoom to fit the viewing area
- The AND operator
- Filter by correlation
- Node labels
- Highlight mode

Creating a window capture and saving it as an image file

You can create a window capture when you want to save a specific connection representation by clicking the **Camera** icon (highlighted in Figure 6-22). Then a window opens where you can select the location to save the image file and select the image file format.



Facet 1 shows: Keywords
Facet 2 shows: Keywords
Node labels: Truncated
Highlight mode: No highlighting
Correlation:

Create a window capture and select the file format

Figure 6-22 Create a window capture and save as image file features

Resuming or pausing rendering

When you select two facets, the Node connections are animated, which takes a while to complete. You can resume or pause the animation in the middle of rendering when you click the **Resume rendering** or **Pause rendering** icons in Figure 6-23.

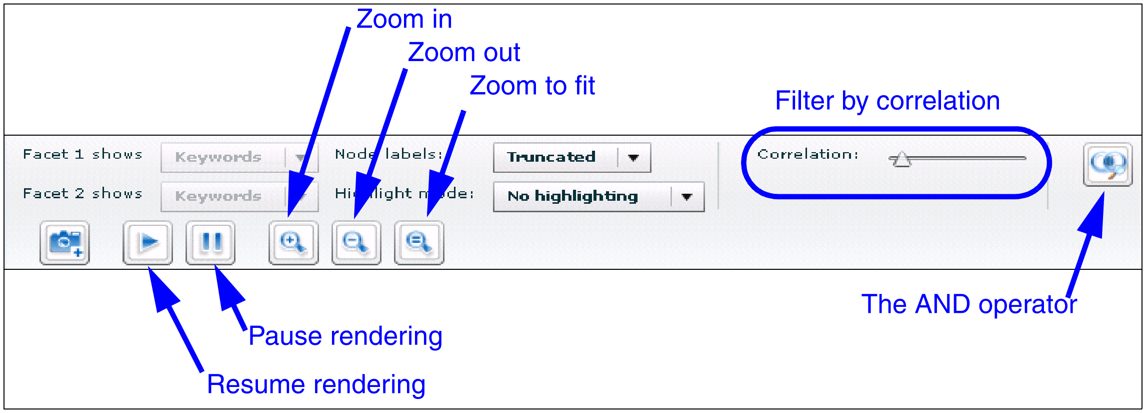

Facet 1 shows: Keywords
Facet 2 shows: Keywords
Node labels: Truncated
Highlight mode: No highlighting
Correlation:

Figure 6-23 Features in the Connections view

Zoom in and Zoom out

Sometimes the Connections view becomes busy when many nodes are to be displayed. You can zoom in and zoom out for specific node connections.

Zoom to fit the viewing area

After you change the browser window size, the Connections view is redrawn if you select the Zoom to fit the viewing area feature. Content Analytics redraws the Connections view to fit the new size of the viewing area.

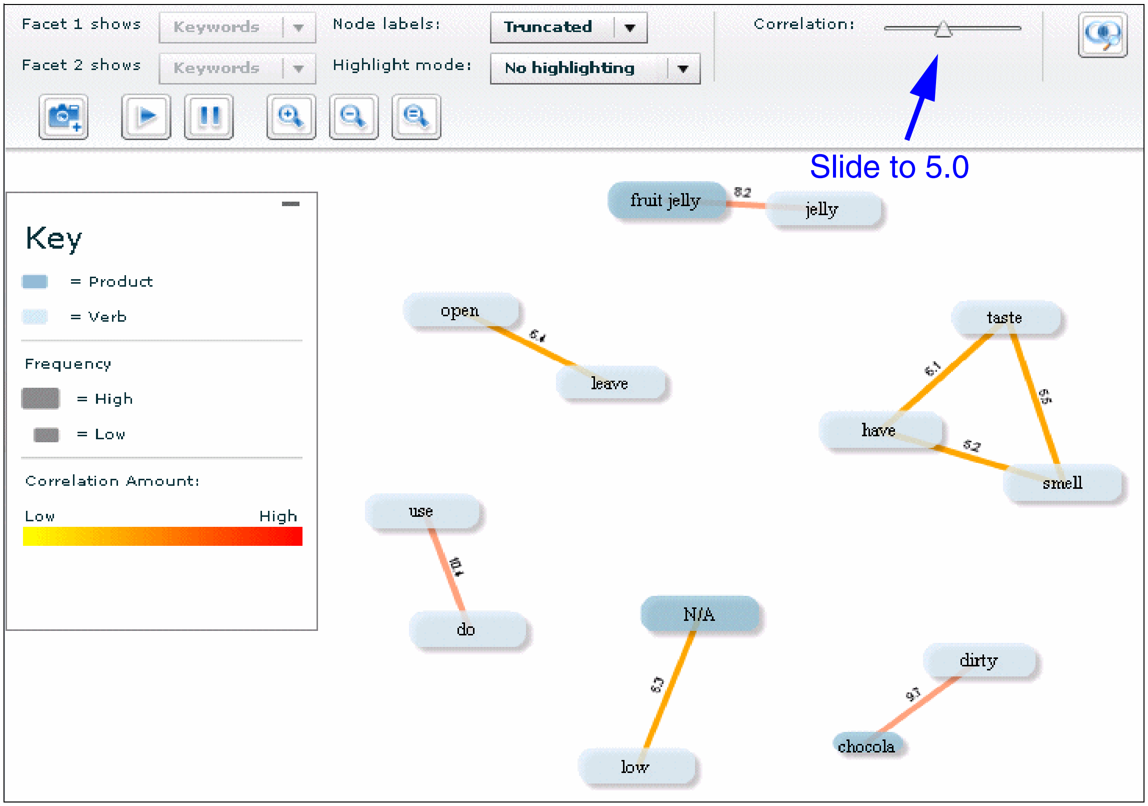
The AND operator

Similar to other views, you can perform the AND operator search to narrow down the scope. The AND operator is helpful in narrowing down the data before interacting with the other views. First, select a specific keyword, and then click the **AND operator** icon.

Filter by correlation

Sometimes the Connection view becomes busy, because, by default, all keywords that have a correlation value greater than 2.0 are displayed. You can set the filter by correlation value to a higher number by using the slide bar. Only keyword pairs that contain a correlation value greater than the correlation value that you set are displayed in the Connection view.

For example, when you want to view the keywords that have a correlation value greater than 5.0, set the filter to 5.0. As a result, fewer nodes are displayed in the Connections view (Figure 6-24) than before the filter change (Figure 6-23 on page 253).



Facet 1 shows: Keywords
Facet 2 shows: Keywords
Node labels: Truncated
Highlight mode: No highlighting
Correlation:

Slide to 5.0

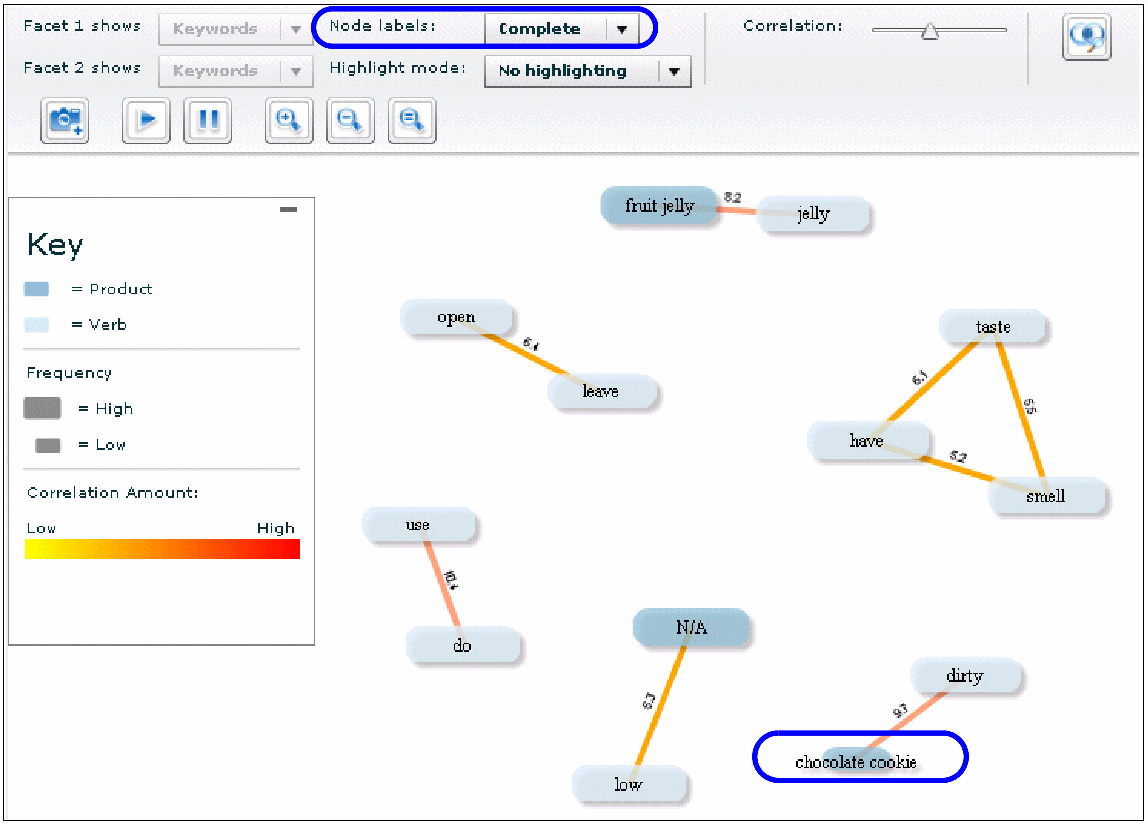
Key
🟦 = Product
🟦 = Verb
Frequency
▇ = High
▆ = Low
Correlation Amount:
Low High

Figure 6-24 Connections view filtered by a correlation value of 5.0

Node labels

You can select how the node label is displayed. The options are Complete, Truncated, or None. By default, the **Truncated** option is selected. To display the full name of the keyword in the node, select the **Complete** option.

For example, the keyword “chocola” and the keyword “dirty” are displayed in the connection view, as shown in Figure 6-24. If you set the Node label field to **Complete**, the keyword “chocola” is changed to “chocolate cookie”, as shown in Figure 6-25.



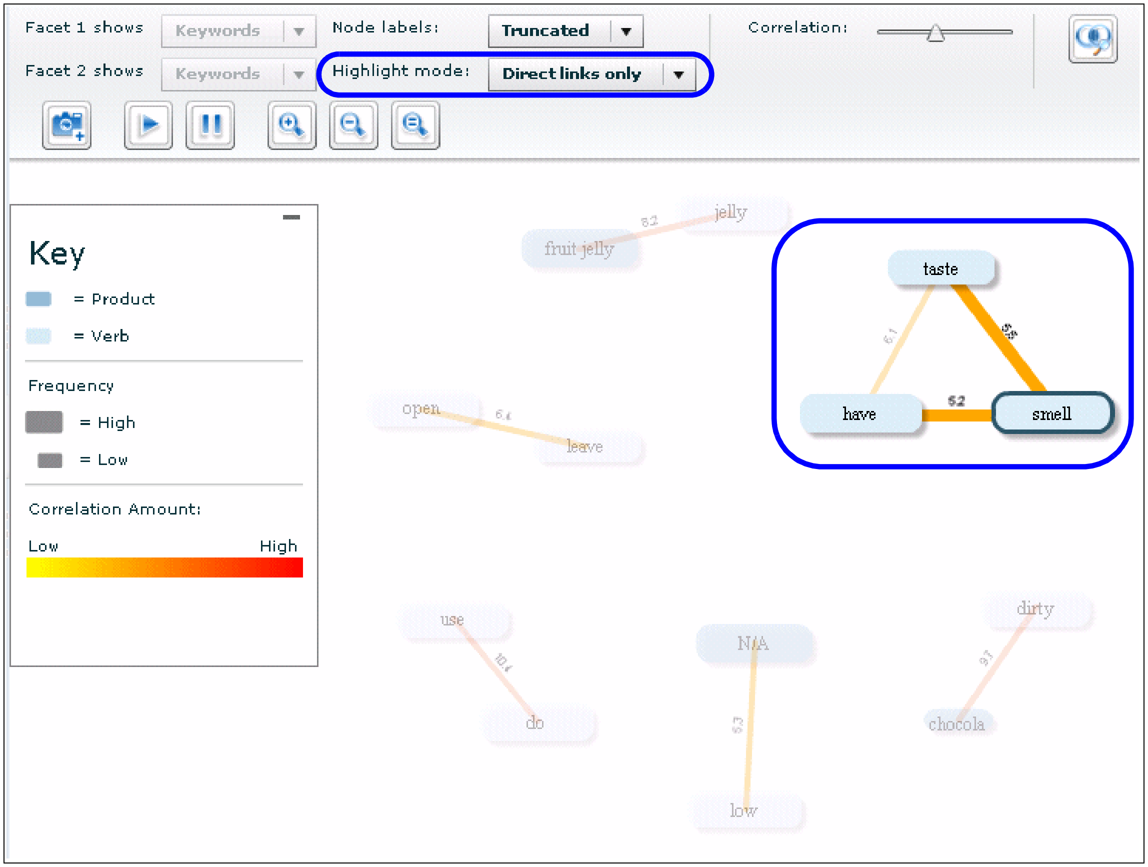
Facet 1 shows: Keywords
Facet 2 shows: Keywords
Node labels: Complete
Highlight mode: No highlighting
Correlation:

Figure 6-25   Selecting Node label as Complete in the Connections view

Highlight mode

Changing the highlight mode is useful when you want to focus on specific keywords in the Connections view. By default, the **No Highlighting** option is selected.

Suppose you focus on the keyword “smell.” To see the direct connection to the keyword, select **Direct links only** for the Highlight mode field. In Figure 6-26, the direct connection to the keyword in the view is now highlighted, and other connections are grayed out.

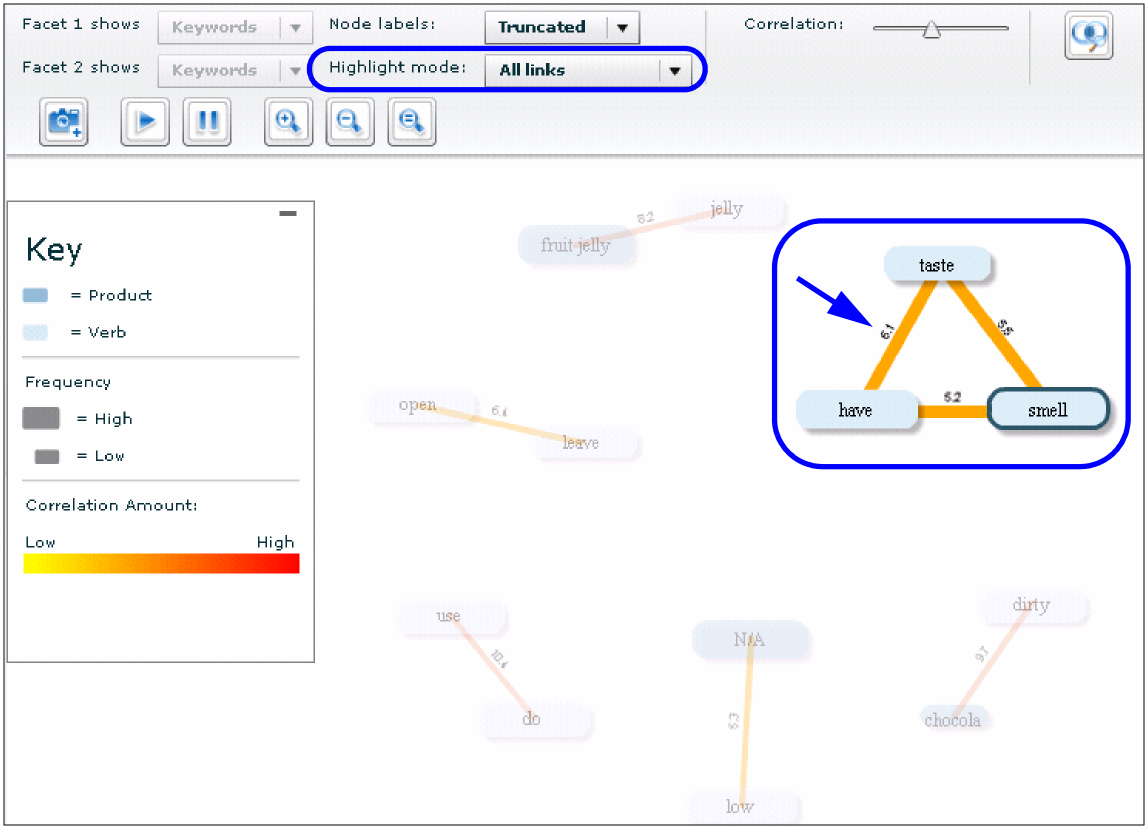


Facet 1 shows: Keywords
Node labels: Truncated
Correlation:
Facet 2 shows: Keywords
Highlight mode: Direct links only

Key
= Product
= Verb
Frequency
= High
= Low
Correlation Amount:
Low High

Figure 6-26 Selecting Direct links only for Highlight mode to view the node “smell”

In addition, when you select **All Links** for the highlight mode field, all the links that are connected to the keywords are highlighted, as shown in Figure 6-27. In our example, the connection between the keyword “have” and the keyword “taste” is now highlighted.



Facet 1 shows: Keywords
Facet 2 shows: Keywords
Node labels: Truncated
Highlight mode: All links
Correlation:

Figure 6-27 Selecting All links for Highlight mode to view the node “smell”

6.8.2 Understanding the Connections view

The Connections view helps you to identify the high correlated keywords based on the selected facets. As explained in the 6.7, “Facet Pairs view” on page 244, Content Analytics requires two sets of search results to calculate the correlation value.

This section helps you to interpret the results of the Connections view and understand how the Connections view is created. In the example in this section, two facets are selected: Product and Verb.

Understanding the values that are displayed

This section explains the meaning of the values that are shown in the Connections view. In Figure 6-28, the keyword “leak” has a higher frequency compared to the other keywords shown in the figure. The node size represents the number of occurrences for the keyword in the result set.

The keywords “orange juice” and “apple juice” have the same blue color, while the keywords “leak” and “drink” have the same light blue color. The color indicates which facet the keywords represent. Thus, you can conclude that the keywords “orange juice” and “apple juice” belong to the same facet (Product). The keywords “leak” and “drink” belong to the same facet (Verb), but the facet is different from the Product facet.

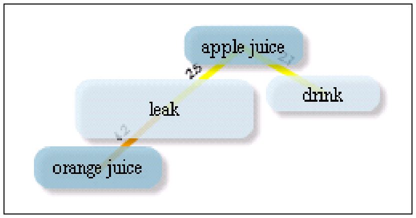


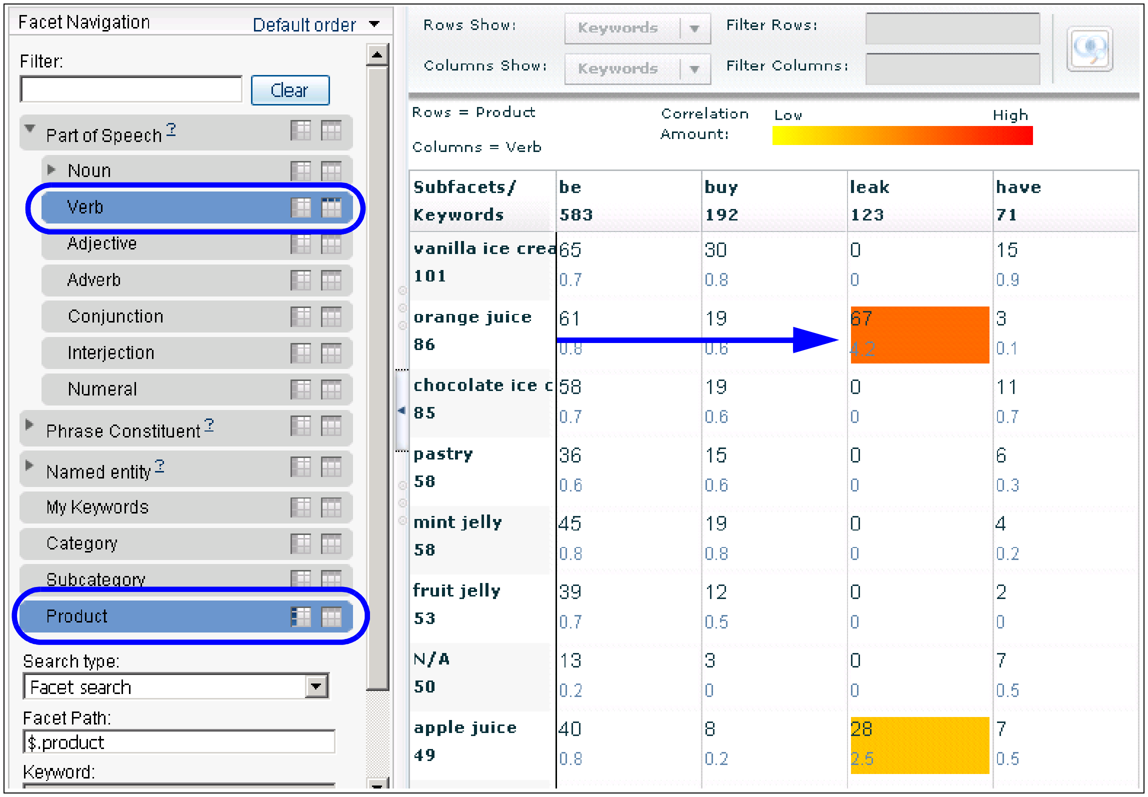
Figure 6-28 Connections view example

When you see the connection between the nodes, you notice that the connection between the keywords “orange juice” and “leak” is an orange color. It is also shorter as compared to the other connections (between the keywords “leak” and “apple juice” or between the keywords “apple juice” and “drink”). Thus, the correlation value between the keywords “orange juice” and “leak” is higher than others. You might want to investigate this relationship further.

How the Connections view is created

As described in the previous section, Content Analytics requires two sets of search results to calculate the correlation value. Content Analytics uses the same calculation mechanism in the Connections view as described in the 6.7, “Facet Pairs view” on page 244. However, it considers all possible combinations of the selected facet pairs.

For example, consider calculating the correlation value between the nodes “orange juice” and “leak” and between the nodes “apple juice” and “leak”, as shown in Figure 6-28. You can see the same correlation value between these keywords when you select the Product facet and the Verb facet in the Facet Pairs view, as shown in Figure 6-29 on page 259. Likewise, the correlation value between the nodes “apple juice” and “drink” is calculated with the same facet pairs, Product facet and Verb facet.



Search type: Facet search

Facet Path: $.product

Keyword:

Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns:

Rows = Product

Columns = Verb

Correlation Amount: Low High

Subfacets/ Keywords	be 583	buy 192	leak 123	have 71
vanilla ice crea 101	65 0.7	30 0.8	0 0	15 0.9
orange juice 86	61 0.8	19 0.6	67 4.2	3 0.1
chocolate ice c 85	58 0.7	19 0.6	0 0	11 0.7
pastry 58	36 0.6	15 0.6	0 0	6 0.3
mint jelly 58	45 0.8	19 0.8	0 0	4 0.2
fruit jelly 53	39 0.7	12 0.5	0 0	2 0
N/A 50	13 0.2	3 0	0 0	7 0.5
apple juice 49	40 0.8	8 0.2	28 2.5	7 0.5

Figure 6-29 Grid view in the Facet Pairs view when selecting Product and Verb

However, the connections sometimes have a different correlation value than Facet Pairs view. The difference in values occurs when keywords from the same facet are compared against each other, such as “smell” and “taste”, “smell” and “have”, “have” and “taste” from the Verb facet, as shown in Figure 6-30.

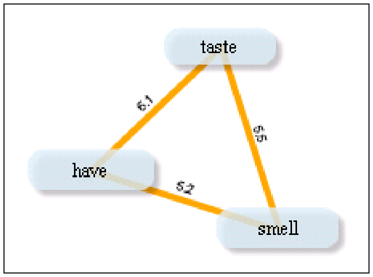
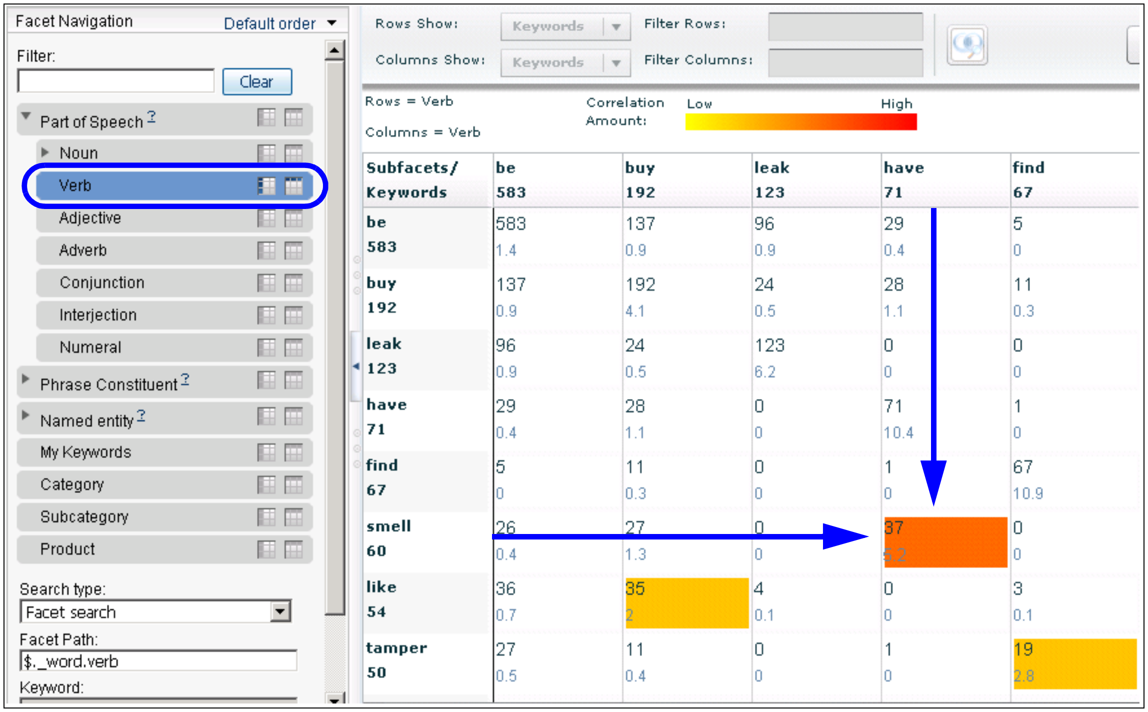


Figure 6-30 Connections view in the same Verb facet

How is the Connections view created in this case? Content Analytics uses the same calculation as the Facet pairs view, but considers all three combinations for the selected facet pairs. In this case, Content Analytics uses the facet pairs, but selects the Verb facet for both vertical facet and horizontal facet, as shown in Figure 6-31.



Facet Navigation Default order

Filter: Clear

- Part of Speech ?
  - Noun
  - Verb
  - Adjective
  - Adverb
  - Conjunction
  - Interjection
  - Numeral
- Phrase Constituent ?
- Named entity ?
- My Keywords
- Category
- Subcategory
- Product

Search type: Facet search

Facet Path: $._word.verb

Keyword:

Rows Show: Keywords Filter Rows:

Columns Show: Keywords Filter Columns:

Rows = Verb

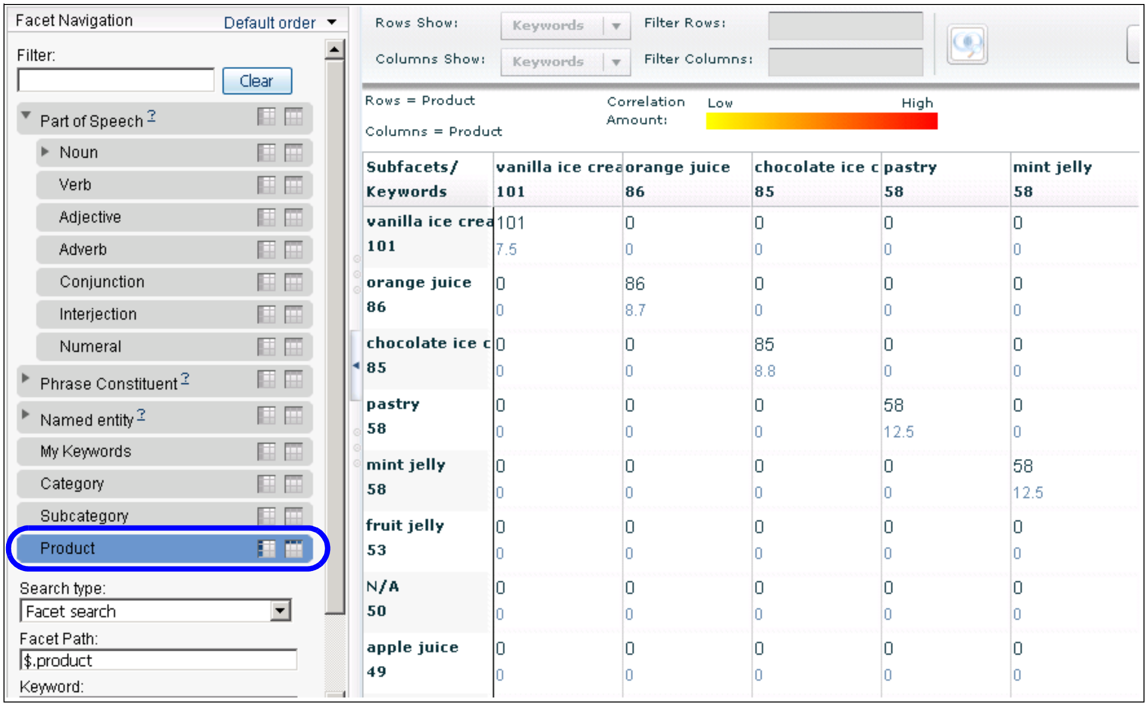
Columns = Verb

Correlation Amount: Low High

Subfacets/ Keywords	be 583	buy 192	leak 123	have 71	find 67
be 583	583 1.4	137 0.9	96 0.9	29 0.4	5 0
buy 192	137 0.9	192 4.1	24 0.5	28 1.1	11 0.3
leak 123	96 0.9	24 0.5	123 6.2	0 0	0 0
have 71	29 0.4	28 1.1	0 0	71 10.4	1 0
find 67	5 0	11 0.3	0 0	1 0	67 10.9
smell 60	26 0.4	27 1.3	0 0	37 5.2	0 0
like 54	36 0.7	35 2	4 0.1	0 0	3 0.1
tamper 50	27 0.5	11 0.4	0 0	1 0	19 2.8

Figure 6-31 Grid view in the Facet Pairs view when selecting Verb and Verb

The remaining combination of the facet pairs in this example relates to selecting the Product facet for both the vertical facet and the horizontal facet. However, as shown in Figure 6-32, these correlation values are not higher than the threshold (2.0 by default). As a result, the connection between these keywords from the Product facet are not displayed.



Search type: Facet search

Facet Path: $.product

Keyword:

Rows Show: Keywords

Columns Show: Keywords

Rows = Product

Columns = Product

Subfacets/ Keywords	vanilla ice crea 101	orange juice 86	chocolate ice c 85	pastry 58	mint jelly 58
vanilla ice crea 101	101 7.5	0 0	0 0	0 0	0 0
orange juice 86	0 0	86 8.7	0 0	0 0	0 0
chocolate ice c 85	0 0	0 0	85 8.8	0 0	0 0
pastry 58	0 0	0 0	0 0	58 12.5	0 0
mint jelly 58	0 0	0 0	0 0	0 0	58 12.5
fruit jelly 53	0 0	0 0	0 0	0 0	0 0
N/A 50	0 0	0 0	0 0	0 0	0 0
apple juice 49	0 0	0 0	0 0	0 0	0 0

Figure 6-32 Grid view in the Facet Pairs view when selecting Product for the horizontal and vertical facet

The Connections view shows the correlation value automatically. You do not need to open the Grid view in the Facet Pairs view and select a different facet each time. You can see how the keywords are correlated with each other immediately so that you can concentrate on highly correlated keywords.

6.8.3 When to use the Connections view

With the Connections view, you can see the keyword cluster based on the correlation. In addition, you can find the hints or "hidden connections" between keywords because the Connections view shows the correlation value for the selected facet pairs and all possible variations of the selected facet pairs.

For example, consider the case when you select the Category facet and the Adjective facet in the Connections view for the sample collection used in this chapter. There is a high correlation between the “strange” keyword and the “Taste/smell” keyword. A high correlation between the “sour” keyword and the “Taste/smell” keyword is displayed. These same high correlation keyword pairs are shown in the Facet Pairs view. However, in the Connections view, you also see the correlation between the “strange” keyword and “sour” keyword. In this case, the three keywords “Taste/smell”, “strange,” and “sour” are connected. Therefore, you can easily see that the user reports a problem when the product taste or smell is sour.

Based on this information, you might drill down into the documents to understand the situation further. When you individually look at the keyword pairs of “sour” and “Taste/smell” or “strange” and “Taste/smell,” you might not think that the user’s report of a “sour” taste as being something unusual. However, in the Connections view, you can easily see the connection through the visual representation of the keyword connections.

You can see the same results in the Facet Pairs view as described earlier. However, you cannot see the correlation of keywords at one time when you use the Facet pairs view. Sometimes you cannot see the keyword pairs with the default configuration in the Grid view (such as “apple juice” and “drink”). To determine this information using the Facet pairs view, you must use the Bird’s eye view which involves extra steps. However, you can find the highly correlated keywords in the Connections view without changing windows.

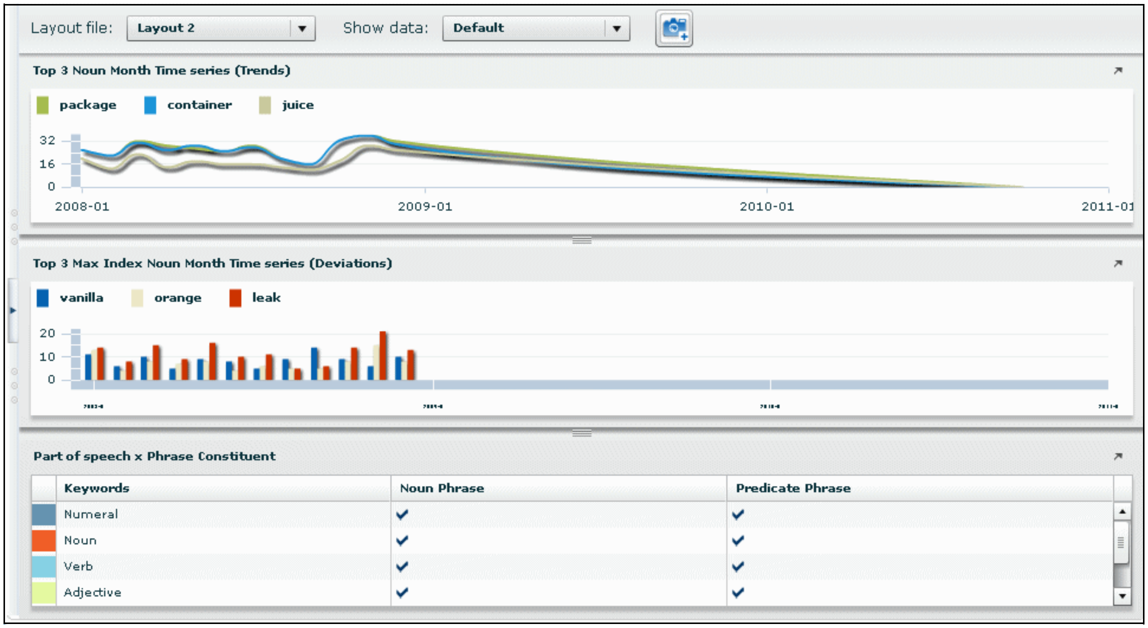
The Connections view shows the “connection” based on the correlation value between all keywords found in the selected facets, and it is easy to filter the result by correlation value. You might see the trends of keyword clusters. As in the Facet Pairs view, after you discover the highly correlated keyword pairs, you can open the Documents view to further investigate the content of the document.

6.9 Dashboard view

The Dashboard view shows various predefined charts and tables in a single text mining view. With this view, you can visualize various aspects of the data to quickly interpret, analyze, and further investigate. In addition, you can save the images in the Bitmap, PNG, or JPEG format so that you can easily share the data with other people for collaboration purposes.

The administrator can customize the predefined Dashboard layouts. Then, the user can select the Dashboard layouts for viewing, saving them as images and analyzing them further. By default, you can use two preconfigured layouts. These default layouts include Layout 1 and Layout 2. The administrator can add additional layouts and customize them based on business requirements.

Figure 6-33 shows an example of the default layout, Layout 2, as it is displayed in the Dashboard.


Layout file: Layout 2

Show data: Default

Top 3 Noun Month Time series (Trends)
package
container
juice
32
16
0
2008-01
2009-01
2010-01
2011-01
Top 3 Max Index Noun Month Time series (Deviations)
vanilla
orange
leak
20
10
0

Part of speech x Phrase Constituent

	Keywords	Noun Phrase	Predicate Phrase
🟦	Numeral	✓	✓
🟧	Noun	✓	✓
🟦	Verb	✓	✓
🟩	Adjective	✓	✓

Figure 6-33 Default Layout 2 in the Dashboard view

6.9.1 Configuring the Dashboard

In this scenario, a new Dashboard layout is configured that contains four separate charts or tables to display data from the collection that created in 4.3.2, “Creating a text analytics collection” on page 90. The new Dashboard layout includes a bar chart, facet table, pie chart, and column chart. In this scenario, we set up each chart to contain separate data.

Creating a layout

The Dashboard layout consists of one or more horizontal or vertical containers. First, you must set up how you want the containers organized in the layout and add panels to each of them. In this scenario, we create a layout that consists of a chart or table in each cell of the layout containing two columns and two rows.

To create a layout, follow the steps:

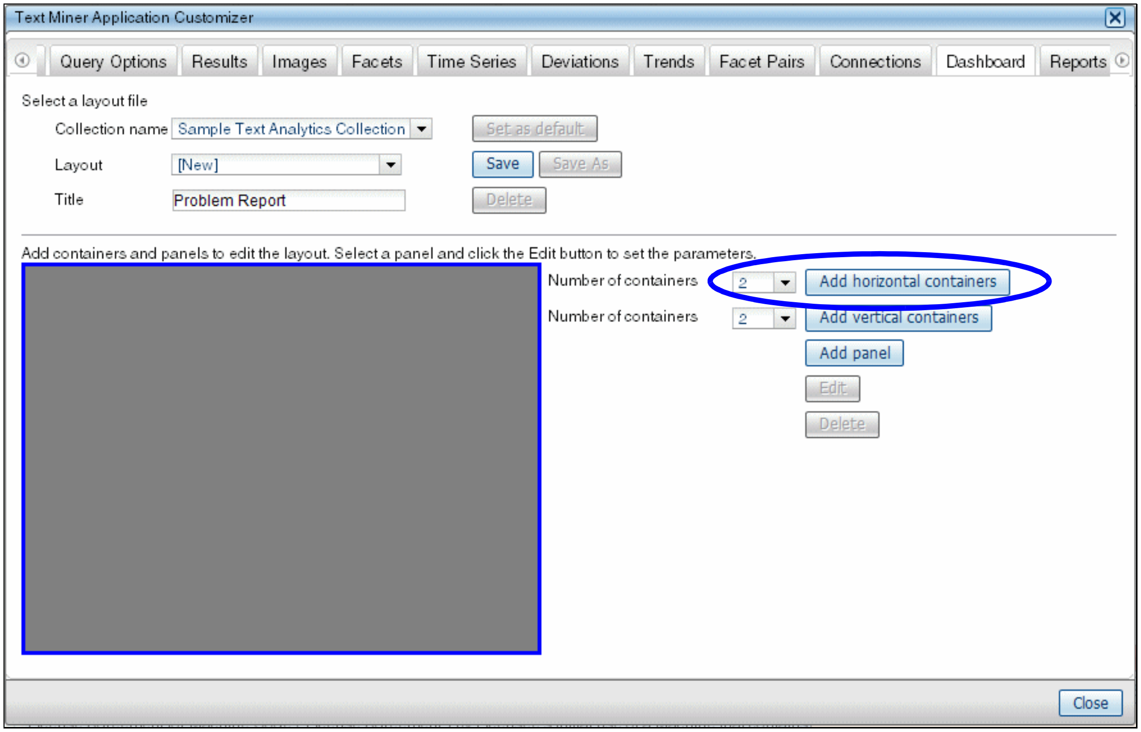
1. In the administration console, click the **Analytics Customizer** tab to open the Analytics Customizer application.
2. Click the **Dashboard** tab.

3. On the Dashboard page (Figure 6-34), complete the following actions:

   a. Select the collection that you want to associate with the new layout in the Collection name field. In this scenario, for Collection name, select **Sample Text Analytics Collection**.

   b. For Layout, select **New**.

   c. For Title, type `Problem Report` as the name for the layout.



Text Miner Application Customizer

Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

Number of containers: 2

Number of containers: 2

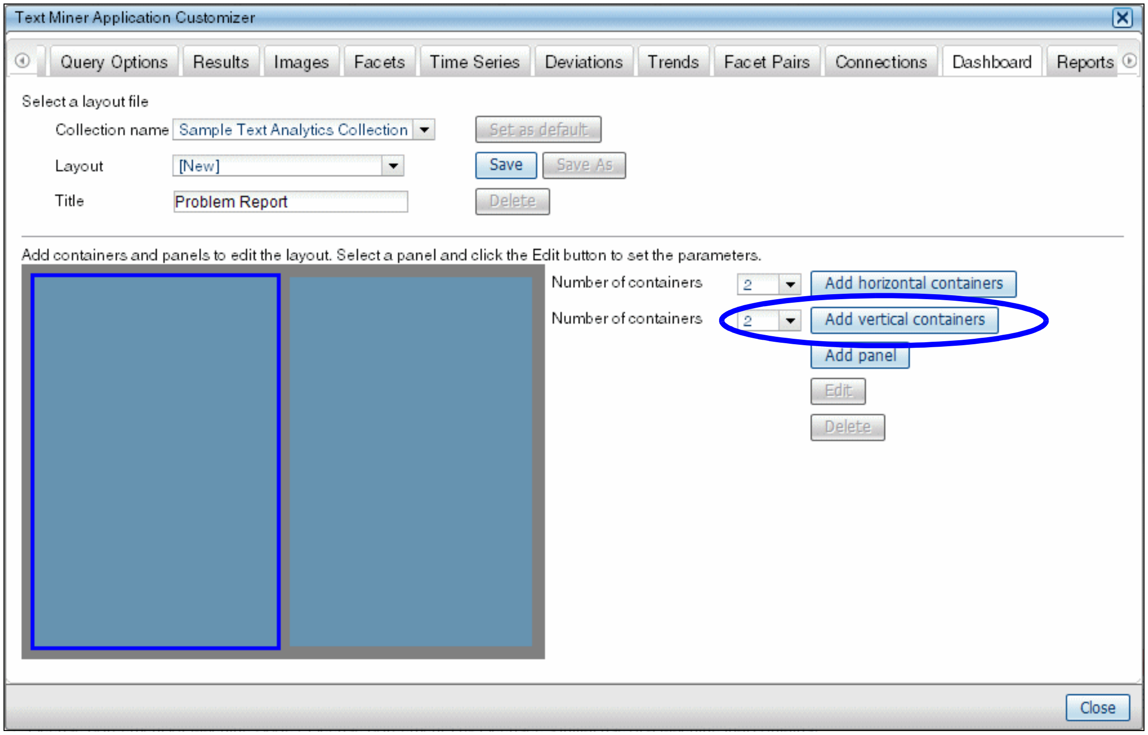
Figure 6-34 Adding two horizontal containers to the new Problem Report Dashboard layout

   The layout consists of containers. Each container needs to contain a panel. If you do not add a panel to each container, you cannot save the new layout.

   d. For Number of containers for the horizontal field, select **2**.

e. Click **Add horizontal container**. You now have two containers of equal size, as shown in Figure 6-35.

f. Click the leftmost horizontal container that you just created to select it, and select **2** for the Number of containers field for the vertical container (Figure 6-35).



Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

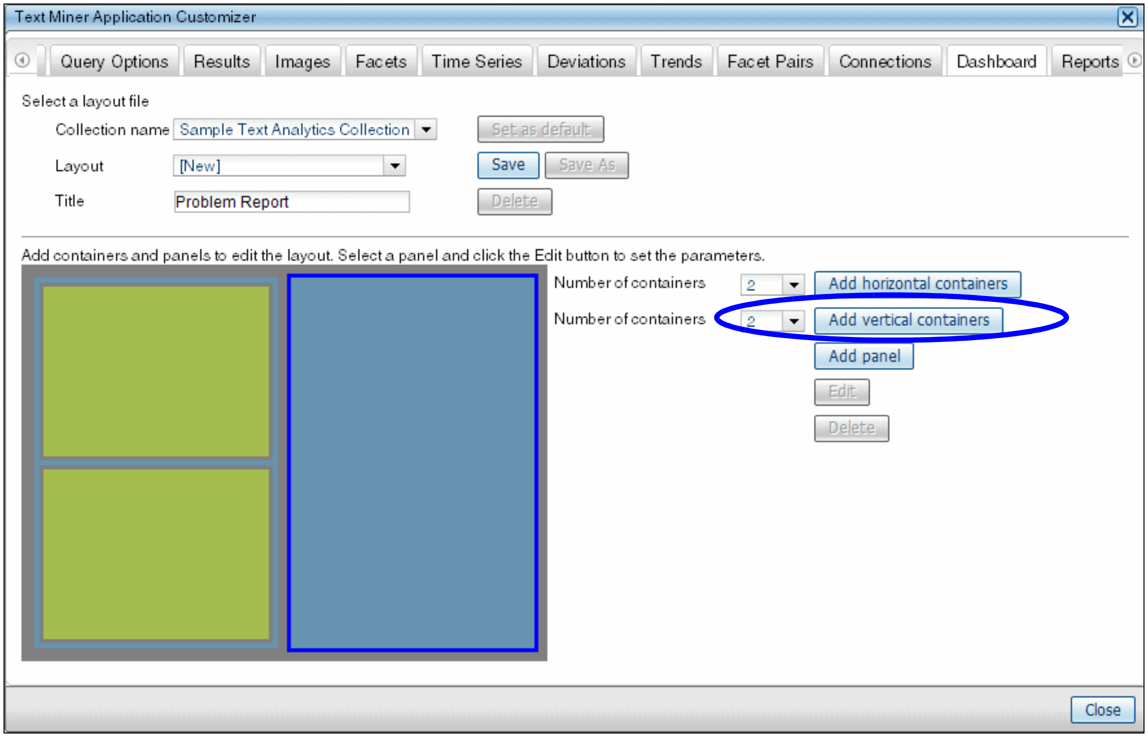
Number of containers: 2

Number of containers: 2

Figure 6-35 Adding two vertical containers to the left container

4. Click **Add vertical container**. You now have three containers. For this scenario, we want to show four charts. Therefore, you need to add one more container to the layout.

5. Click the rightmost horizontal container to select it, and select **2** for the Number of containers for the vertical container, as shown in Figure 6-36.



Text Miner Application Customizer

Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

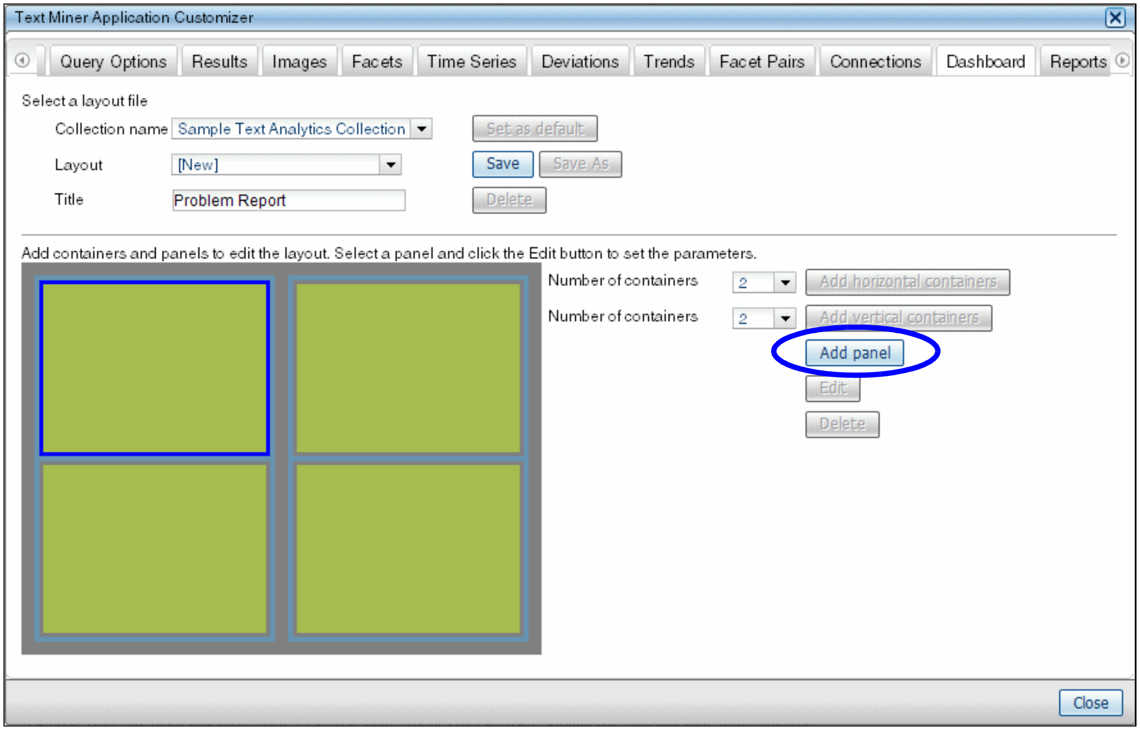
Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

Number of containers: 2

Number of containers: 2

Figure 6-36 Adding two vertical containers to the right container to create a four container layout

6. Click **Add vertical container**. You now have four containers of equal size, as shown in Figure 6-37.



Text Miner Application Customizer

Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

Number of containers: 2

Number of containers: 2

Figure 6-37 Adding a panel to the upper left container of the new layout

7. Click the upper-left container, and click **Add Panel**.
8. Repeat step 7 for each container so that each of the four containers contains a panel. This action adds a panel to each container with the default settings for a bar chart.

The following sections explain how you can edit the panel to create your specific charts and tables.

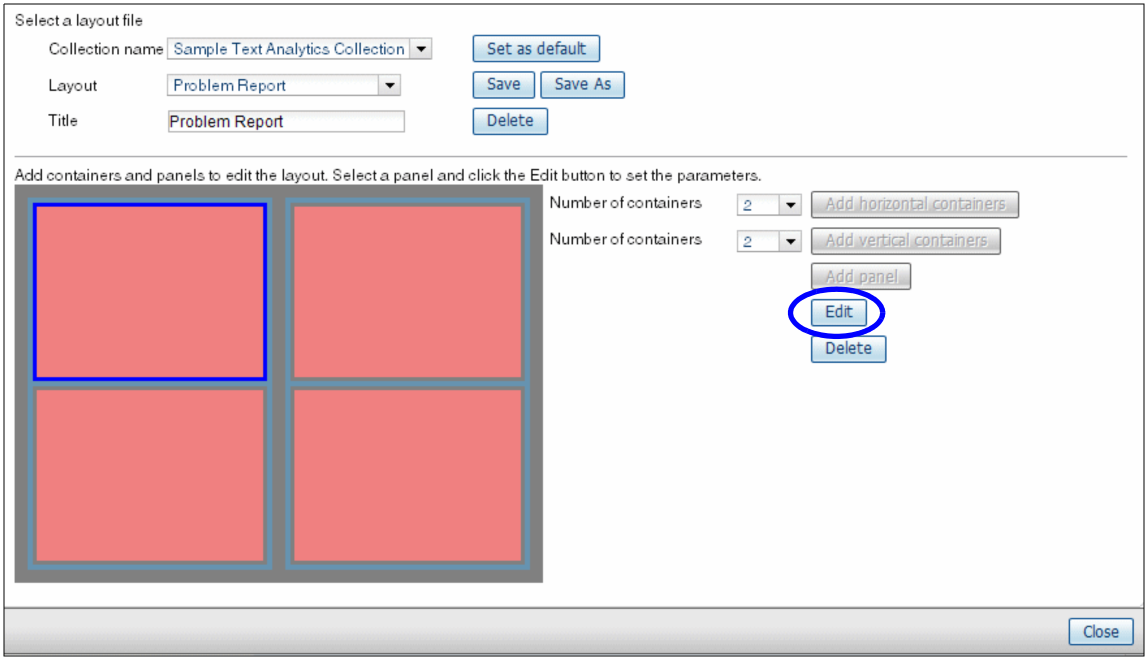
Creating a bar chart

In this section, you configure the new layout to contain a bar chart located in the upper-left panel of the new layout view. You set the data to be displayed in the bar chart to be the frequency of the top category facet values. For more information about the Dashboard configuration options, see the following address:

http://www-01.ibm.com/support/docview.wss?uid=swg21420024

To create the bar chart, follow these steps:

1. Click the upper-left panel to select it and click **Edit** (Figure 6-38).



Select a layout file

Collection name: Sample Text Analytics Collection

Layout: Problem Report

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

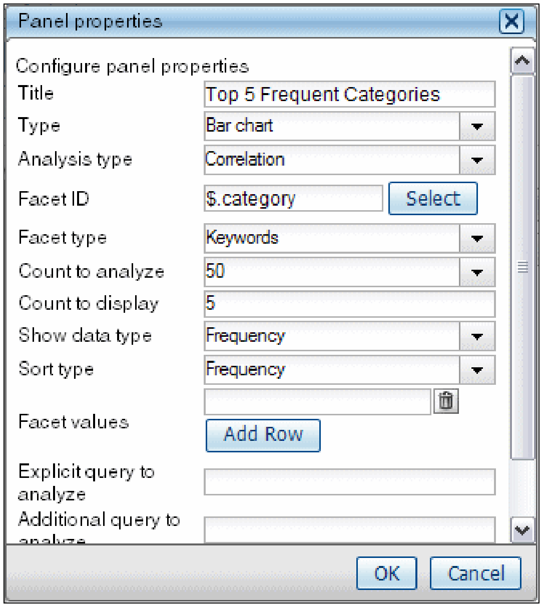
Number of containers: 2

Number of containers: 2

Figure 6-38 Dashboard layout panel to configure the bar chart

2. In the Panel properties window (Figure 6-39), complete the following steps:

   a. For the Title field, type `Top 5 Frequent Categories`.
   b. For the Type field, select **Bar chart**.
   c. For the Facet ID field, click **Select** and click **Category**.
   d. Click **OK**.



Panel properties

Configure panel properties

Title	Top 5 Frequent Categories
Type	Bar chart
Analysis type	Correlation
Facet ID	$.category
Facet type	Keywords
Count to analyze	50
Count to display	5
Show data type	Frequency
Sort type	Frequency
Facet values	
Explicit query to analyze	
Additional query to analyze	

Figure 6-39 The bar chart panel properties

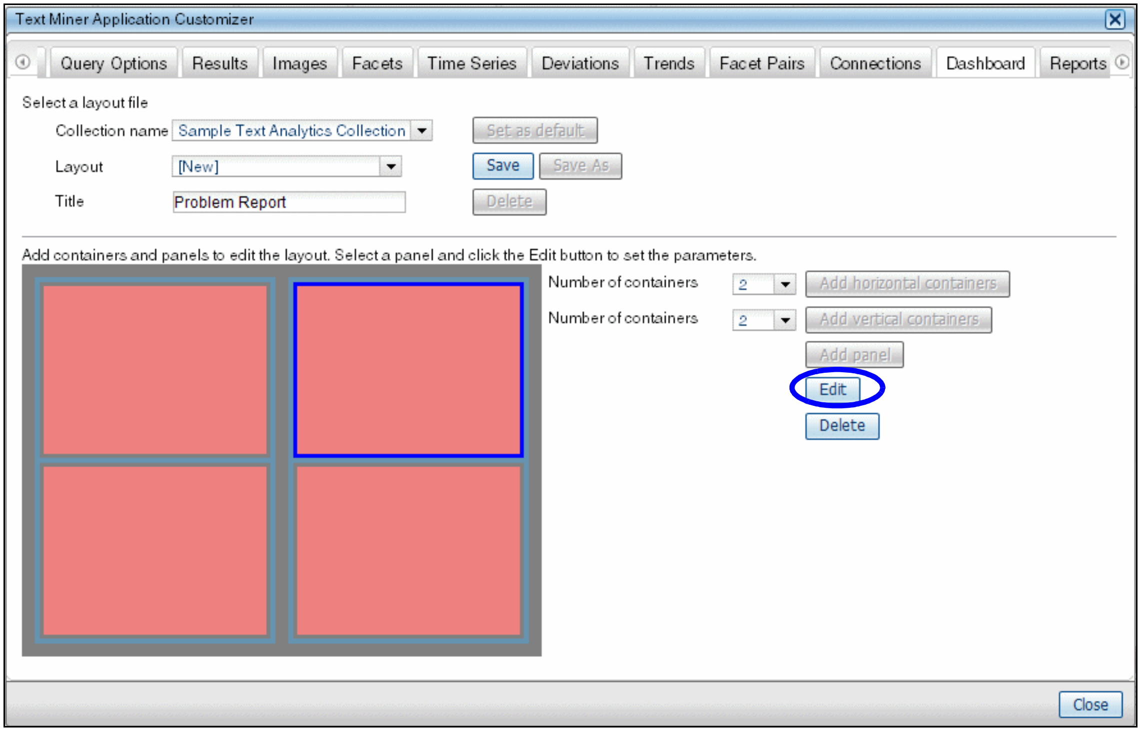
You do not see the results of the bar chart until all containers are configured. For an example of the completed bart chart, see the upper-left container in Figure 6-43 on page 275.

Creating a facet table

In this section, you configure the new layout to include a facet table located in the upper-right panel of the layout view. You set the data to be displayed in the facet table as the top correlated verbs.

To create a facet table, follow these steps:

1. Click the upper-right panel to select it and click **Edit** (Figure 6-40).



Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

Number of containers: 2

Number of containers: 2

Figure 6-40 Dashboard layout panel to configure the facet table

2. Add the field values as shown in Table 6-1. Keep the default values for all other fields.

Table 6-1 Panel properties for the top 5 correlated verb facets

Field	**Value**
Title	Top 5 Correlated Verbs
Type	Facet Table
Facet ID	Select **Part of Speech** → **Verb**
Show data type	Correlation or index
Sort type	Correlation

3. Click **OK**.

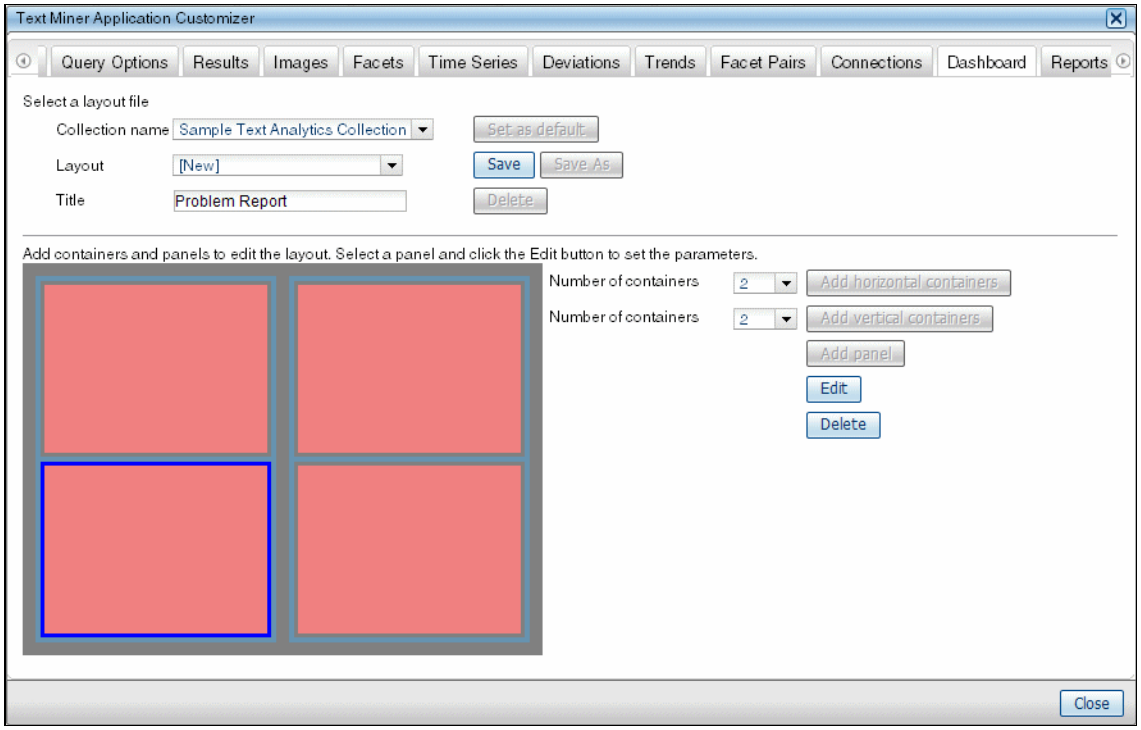
You do not see the results of the facet table until all containers are configured. For an example of the completed facet table, see the upper-right container in Figure 6-43 on page 275.

Creating a pie chart

In this section, you configure the new layout to contain a pie chart located in the lower left panel of the layout view. You set the data to be displayed in the pie chart as the correlated values of the Product facet.

To create a pie chart, follow these steps:

1. Click the lower left panel to select it and click **Edit** (Figure 6-41).



Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

Number of containers: 2

Number of containers: 2

Figure 6-41 Dashboard layout panel to configure for the pie chart

2. Add the field values as shown in Table 6-2 and keep the default values for all other fields.

Table 6-2 Panel properties for the top 5 correlated products pie chart

Field	**Value**
Title	Top 5 Products by Correlation

Field	Value
Type	Pie Chart
Facet ID	Select **Product**
Show data type	Correlation or index
Sort type	Correlation

3. Click **OK**.

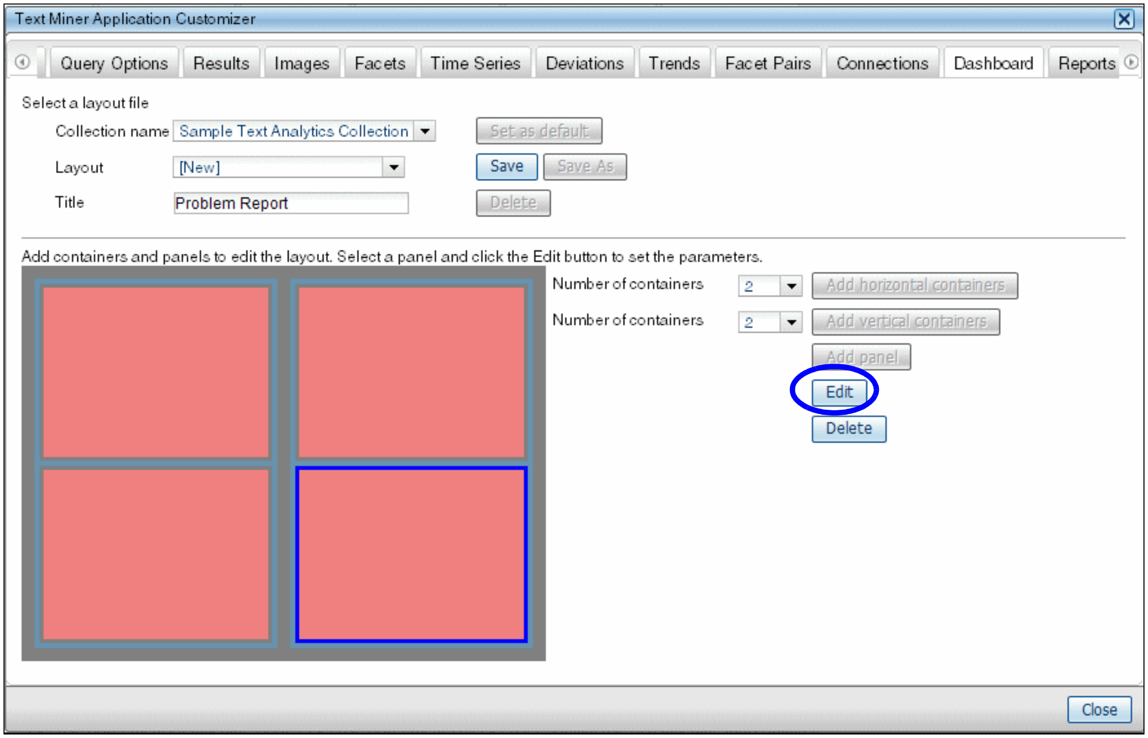
You do not see the results of the pie chart until all containers are configured. For an example of the completed pie chart, see the lower-left container in Figure 6-43 on page 275.

Creating a column chart

You configure the fourth panel to contain a column chart located in the lower right panel of the layout view. You set the data to be displayed in the column chart to the frequent values of the subcategory facet for documents that contain the term “juice”. The data set is narrowed to documents that contain the term “juice.” Then the top 50 most frequent subcategory values are analyzed, and the top three subcategories are displayed in the chart.

To create a column chart, follow these steps:

1. Click the lower right panel to select it and click **Edit** (Figure 6-42).



Text Miner Application Customizer

Select a layout file

Collection name: Sample Text Analytics Collection

Layout: [New]

Title: Problem Report

Add containers and panels to edit the layout. Select a panel and click the Edit button to set the parameters.

Number of containers: 2

Number of containers: 2

Figure 6-42 Dashboard layout panel to configure for the column chart

2. Add the field values as shown in Table 6-3 and keep the default values for all other fields.

Table 6-3 Panel properties for the top three frequent subcategories column

Field	**Value**
Title	Top 3 Frequent Subcategories Containing Juice
Type	Column Chart
Facet ID	Select **Subcategory**
Count to display	3
Explicit query to analyze	juice

3. Click **OK**.

You do not see the results of the column chart until all containers are configured. For an example of the completed column chart, see the lower-right container in Figure 6-43 on page 275.

Saving the Dashboard layout

Now that you configured the new Dashboard layout, you need to save it:

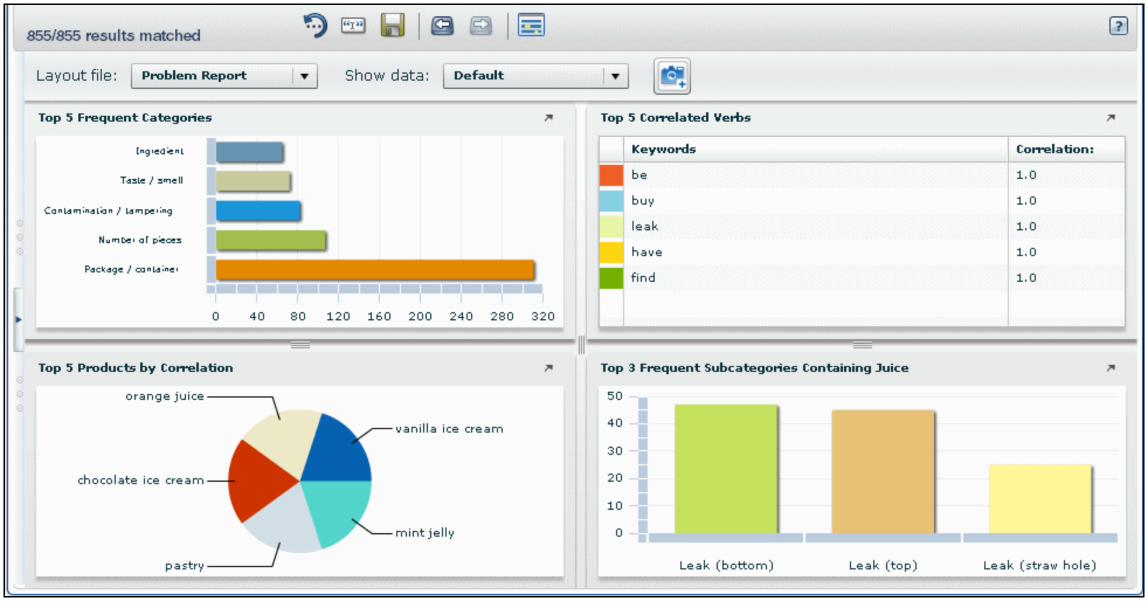
1. Click **Save**.
2. Click **Close** to close the Dashboard configuration window.
3. Click **Save Changes**. Otherwise, the changes might not fully save to the text miner application.
4. Click **Exit** to exit the Analytics Customizer application.
5. In the message window that indicates that the window will close, click **OK**.

6.9.2 Viewing the Dashboard

After an administrator sets up a new dashboard layout, a user can view it and work with it through the text miner application:

1. Open the text miner application.
2. For this scenario, use the **Sample Text Analytics Collection**. If this collection is not displayed at the top of the text miner application, click the **change** link, select the **Sample Text Analytics Collection**, and click **Save**.
3. Click the **Dashboard** tab. In the Layout File field, select the **Problem Report** layout.

The layout that you created in 6.9.1, "Configuring the Dashboard" on page 263, is now shown in the window. The four charts and tables contain the analytic data of the collection, as shown in Figure 6-43. If a data value is listed in more than one chart or table, the color associated with that data value is the same across the multiple charts and tables.



Layout file: Problem Report

Show data: Default

Top 5 Correlated Verbs

	Keywords	Correlation:
🟧	be	1.0
🟦	buy	1.0
🟨	leak	1.0
🟨	have	1.0
🟩	find	1.0

Figure 6-43 Problem Report layout in the Dashboard view

6.9.3 Working with the Dashboard

The Dashboard provides additional useful functionality. For example, you can further narrow down the data set, display correlation and frequency values for a data point, enlarge a table or chart, and set user preferences.

Narrowing the search results

With the charts and tables in the Dashboard layout, you can further narrow down your data set to focus on an area of interest. In this scenario, you click the **Package/container** bar in the "Top 5 Frequent Categories" chart. As a result, the following syntax is automatically added to the query statement:

```
/"keyword$.category"/"Package / container"
```

Notice that all the charts and tables have changed, except for the column chart. Also only the documents that contain a category equal to Package/container are now displayed. Because the column chart has the term "juice" set as the explicit query to analyze, it is not modified by any additional query statements that a user includes. If you want the user's query statement to be appended to the query defined for the chart, set the query in the additional query to analyze the field when editing the panel in the dashboard layout.

To view the query syntax area, click the **Expand this Area** icon at the top of the window. Figure 6-44 on page 277 shows the addition to the query syntax and the updated charts and tables.

Frequency and correlation display

When you move the mouse pointer over a data point on the chart or table, you see the frequency and correlation values for that particular data point. In this scenario (Figure 6-44 on page 277), move the mouse over the **Mineral** pie piece in the "Top 5 Products by Correlation" pie chart to view the frequency of 39 and correlation of 1.9.

Expanding and minimizing a chart

To see only one chart at a time in a larger view, click the **expand** icon in the upper-right corner of the chart or table (Figure 6-44 on page 277).

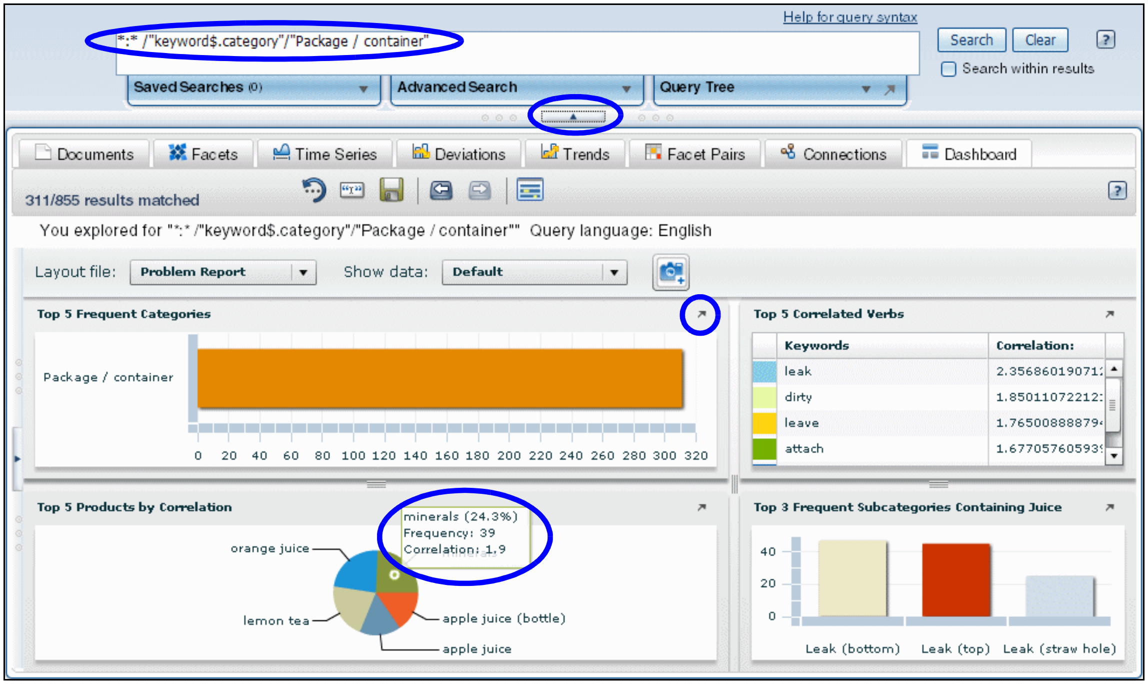
After the chart is expanded, click the **minimize** icon in the upper right corner of the chart or table to go back to the main layout view that shows all of the charts or tables.

Changing the chart or table size

To change the size of the charts, you move the mouse pointer over the border of the chart, and drag the frame of the chart to your desired location. With this action, you keep all the charts in the layout viewable while making a chart larger. Figure 6-44 on page 277 shows an example of changing the frame of a chart. Notice that the "Top 5 Frequent Categories" and "Top 5 Products by Correlation" charts are wider than the other charts and tables.

Dashboard preference settings

To change the default layout that is shown in the Dashboard, select **Preferences** → **Dashboard**. For the Sample Text Analytics Collection, select **Problem Report** to be the default layout. Now the Problem Report layout is displayed every time you open the Dashboard view for the Sample Text Analytics Collection.



: /"keyword$.category"/"Package / container"

311/855 results matched

You explored for "*:* /"keyword$.category"/"Package / container"" Query language: English

Layout file: Problem Report

Show data: Default

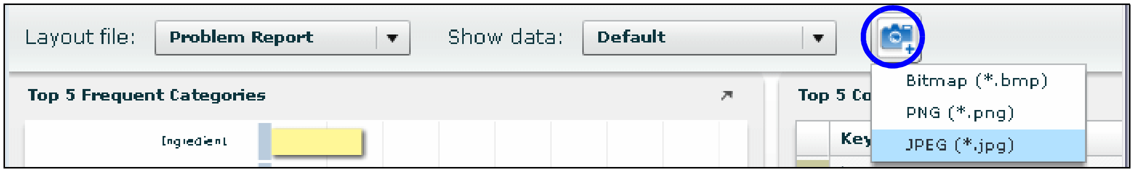
Top 5 Correlated Verbs

	Keywords	Correlation:
	leak	2.35686019071
	dirty	1.85011072212
	leave	1.76500888879
	attach	1.67705760593

Figure 6-44 Problem Report layout with results for the Package/container category

6.9.4 Saving Dashboard charts as images

With the Dashboard, you can save all of the charts and tables in the selected layout as an image, or you can save an individual expanded chart or table. You can use the Bitmap, PNG, and JPEG formats to save the images. You save an image by clicking the **image** icon, as shown in Figure 6-45, and selecting your desired image format. After the image is saved, you can share it with coworkers for further collaboration.



Layout file: Problem Report

Show data: Default

Figure 6-45 Saving the Dashboard charts by clicking the image icon

7

Performing content analysis

Chapter 5, “Text miner application: Basic features” on page 143, provides information about the basic features and operations of the text miner application. Among them, are the search and discovery features and the views that are available for analysis. This chapter focuses on performing content analysis with the text miner application by using use-case scenarios to discover actionable insight from textual data.

This chapter includes the following sections:

- Discovering actionable insight with the text miner application
- Content analysis scenarios
- Configuring the Dictionary Lookup annotator
- Configuring the Pattern Matcher annotator
- Preferred practices

7.1 Discovering actionable insight with the text miner application

The text miner application, which is packaged with IBM Content Analytics, is a powerful tool with which you can analyze your data in many ways and that is ready for immediate use. This section explains how to use the text miner application to discover actionable insight from your textual data.

Throughout this section, the Sample Text Analytics Collection is used that is created when you select **Text Analytics Tutorial** in the First Steps program. Create this collection before you proceed with the rest of the section. You can manually build the text analytics collection with the same data and configuration as explained in Chapter 4, “Installing and configuring IBM Content Analytics” on page 71.

This section also explains how to define a custom dictionary and custom text analysis rules. For the specific steps to create a custom dictionary and custom text analysis rules, see 7.3, “Configuring the Dictionary Lookup annotator” on page 299, and 7.4, “Configuring the Pattern Matcher annotator” on page 309.

7.1.1 The sample data

The data used in this example is in the `ES_INSTALL_ROOT/samples/firststep/data/xml/xmls.tar.gz` file. When you extract each XML data file from the `.tar` file, you see the content similar to what is shown in Example 7-1.

Example 7-1 XML data used in the Sample Text Analytics Collection

```
<?xml version="1.0" encoding="UTF-8"?>
<doc>
    <id>00000000</id>
    <title>lemon tea - Package / container</title>
    <date>2008-01-01</date>
    <timestamp>1199186392296</timestamp>
    <category>Package / container</category>
    <subcategory>Straw</subcategory>
    <product>lemon tea</product>
    <text>[Pack] The straw was peeled off from the juice pack.</text>
</doc>
```

This data simulates the transcript of inbound customer calls made to a call center data of Fictitious Confectionery Company A.

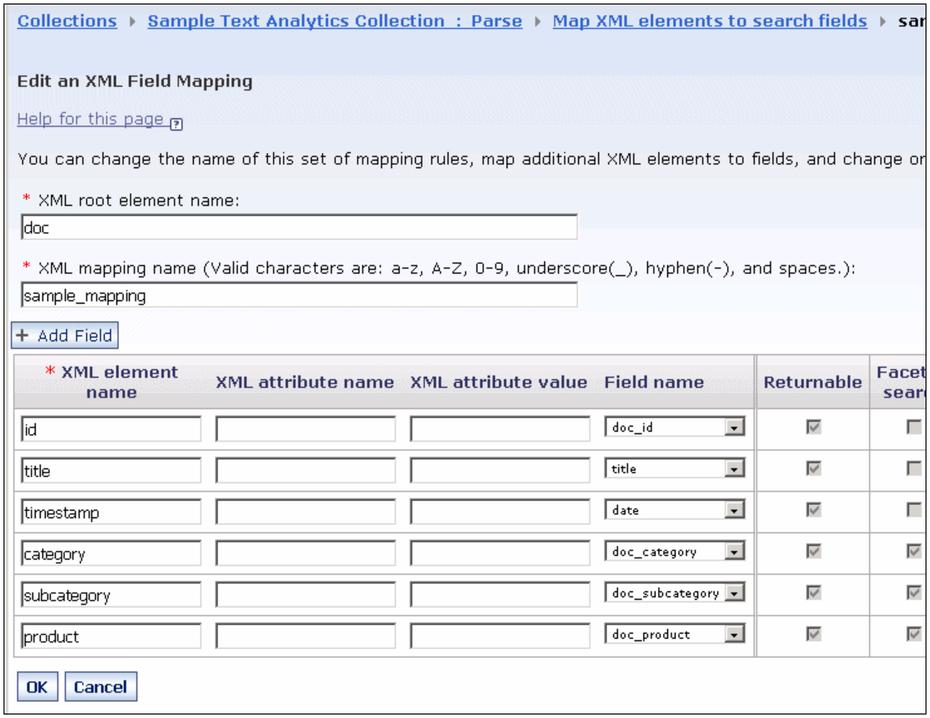
Sample data source: The sample data is bundled with IBM Content Analytics.

When you run the Text Analytics Tutorial from the First Steps program, the Sample Text Analytics Collection is created for you with the sample data. When the Sample Text Analytics Collection is created, a file system crawler (depending on the Content Analytics server platform) is created automatically, and the XML elements are mapped to the appropriate facets. The Sample Text Analytics Collection is configured with predefined facets, but no dictionary and custom text analysis rules have yet been defined.

Consequently, at this point, you have a text analytics collection with analyzed data but without a dictionary or custom text analysis rules. However, the Text Analytics Tutorial creates the following mapping rules for you:

- XML elements to the search fields
- A facet to the field mapping

The predefined facets in the Facet Navigation pane on the left side of the text miner application. Figure 7-1 shows the mapping named *sample_mapping* that defines the mapping between the XML element name and Field name.


Collections ▸ Sample Text Analytics Collection : Parse ▸ Map XML elements to search fields ▸ sar

Edit an XML Field Mapping

You can change the name of this set of mapping rules, map additional XML elements to fields, and change or

* XML root element name: doc

* XML mapping name (Valid characters are: a-z, A-Z, 0-9, underscore(_), hyphen(-), and spaces.): sample_mapping

* XML element name	XML attribute name	XML attribute value	Field name	Returnable	Facet sear
id			doc_id	☑	□
title			title	☑	□
timestamp			date	☑	□
category			doc_category	☑	☑
subcategory			doc_subcategory	☑	☑
product			doc_product	☑	☑

Figure 7-1 Mapping the XMl element to the search field for sample data

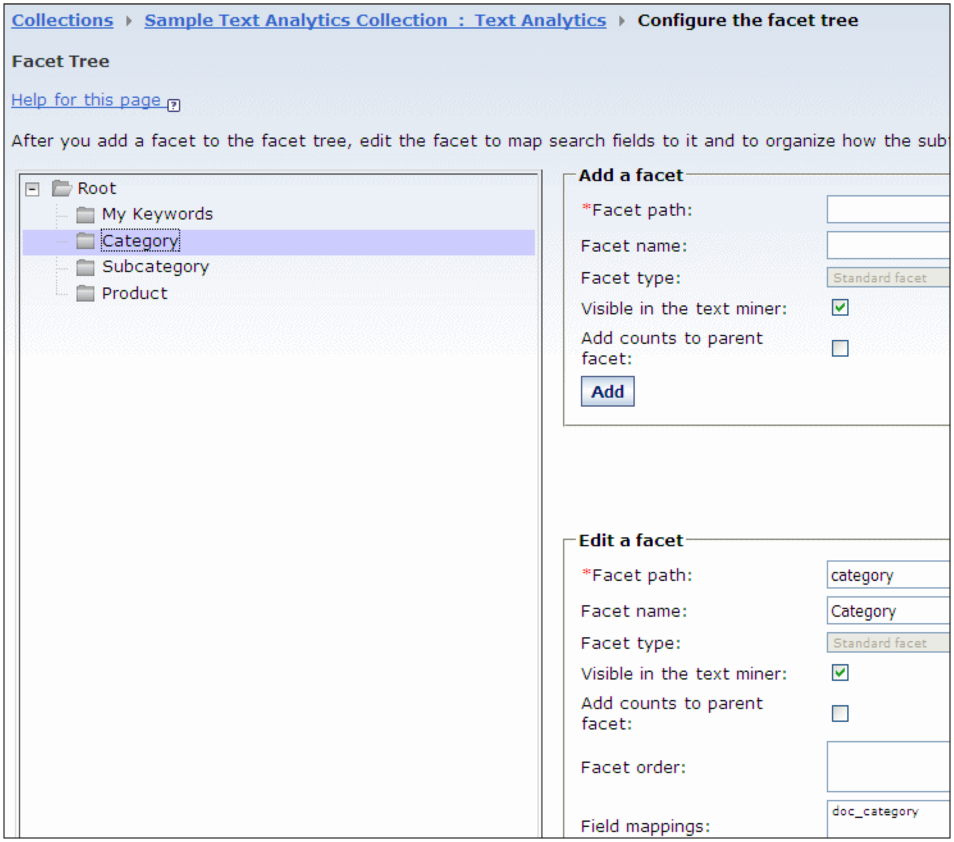
Figure 7-1 on page 281 also shows that the sample XML document has the element and field associations shown in Table .

Element and field associations for the sample XML document

Element	Field
id	doc_id
title	title
timestamp	date
category	doc_category
subcategory	doc_subcategory
product	doc_product

Figure 7-2 on page 283 shows the facets that are created for you. The mapping between facets and search fields is automatically done when you run the Text Analytics Tutorial. Three default facets are created by the tutorial: the Category facet, the Subcategory facet, and the Product facet. These facets are mapped with the following fields:

- The Category facet is mapped with the doc_category field.
- The Subcategory facet is mapped with the doc_subcategory field.
- The Product facet is mapped with the doc_product field.


Collections ▸ Sample Text Analytics Collection : Text Analytics ▸ Configure the facet tree

Facet Tree

Help for this page

After you add a facet to the facet tree, edit the facet to map search fields to it and to organize how the sub

Add a facet

*Facet path:

Facet name:

Facet type: Standard facet

Visible in the text miner: ☑

Add counts to parent facet: □

Edit a facet

*Facet path: category

Facet name: Category

Facet type: Standard facet

Visible in the text miner: ☑

Add counts to parent facet: □

Facet order:

Field mappings: doc_category

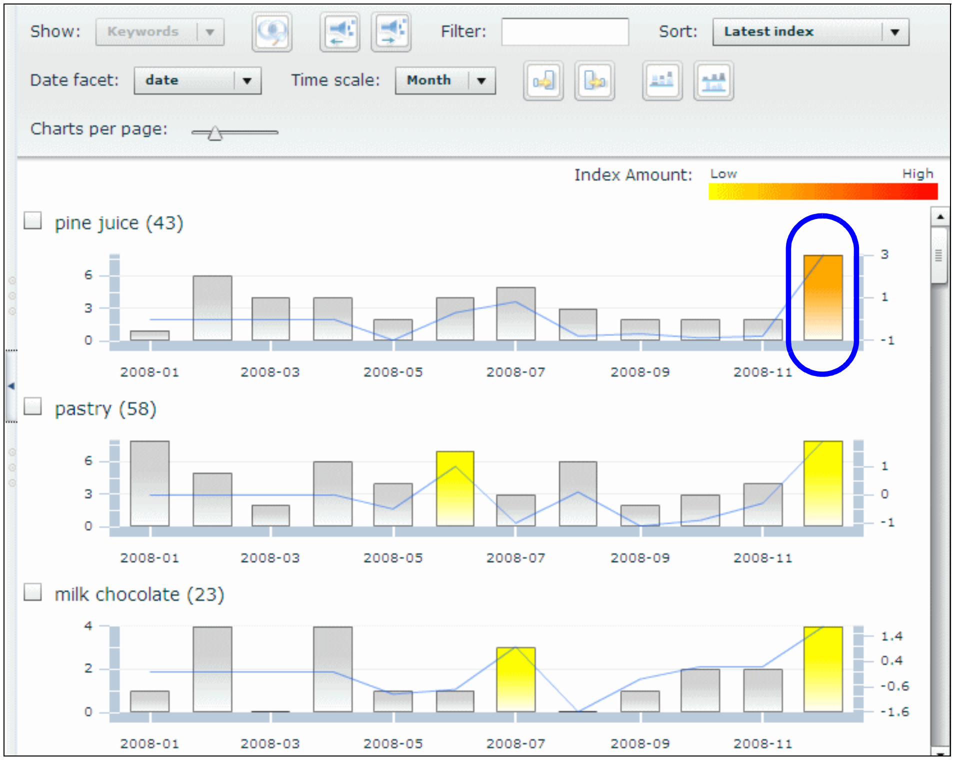
Figure 7-2 Mapping between the facets and search fields

Notice that the text XML element in Example 7-1 on page 280 is *not* associated with any search field. The content in the text XML element is intentionally not associated with a facet.

7.1.2 Insights without customization

The text analytics collection is ready for analysis without any custom dictionary or custom text analysis rules. You can derive many insights from the sample collection.

For example, in the Trends view, you can tell if a particular facet has increased in numbers over a period. Figure 7-3 shows the result when you select the Product facet and sort it by the latest index. Notice that the number of calls (as indicated by the number of cases in the sample data) related to pine juice increases in December 2008 (highlighted in Figure 7-3).



Show: Keywords

Filter:

Sort: Latest index

Date facet: date

Time scale: Month

Charts per page:

Index Amount: Low High

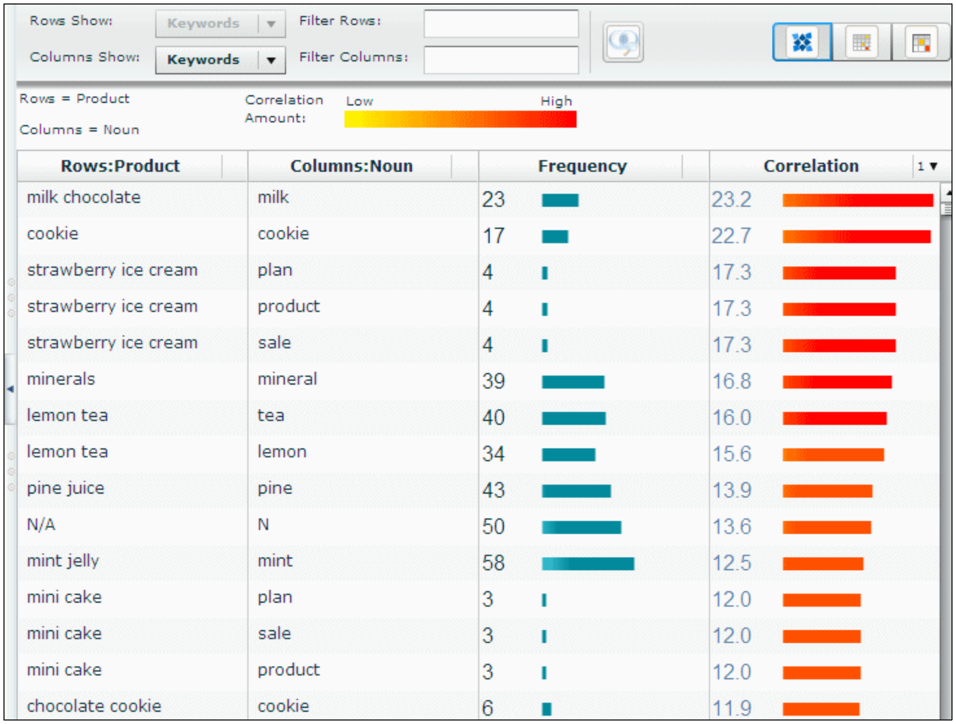
□ pine juice (43)

□ pastry (58)

□ milk chocolate (23)

Figure 7-3 Trends view showing the Product facet sorted by the latest index

In the Facet Pairs view, you can also see which keywords are highly correlated. Figure 7-4 shows the result when you select the Product facet and the Noun facet and sort them by correlation.



Rows Show: Keywords
Columns Show: Keywords
Filter Rows:
Filter Columns:

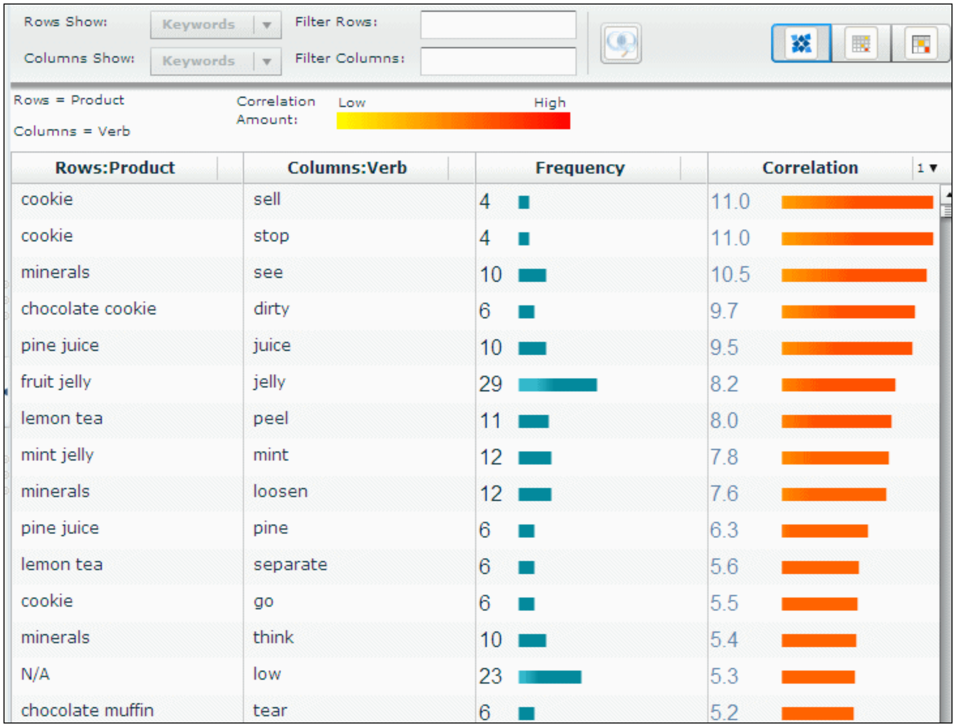
Rows = Product
Columns = Noun

Correlation Amount: Low High

Rows:Product	Columns:Noun	Frequency	Correlation 1 ▼
milk chocolate	milk	23	23.2
cookie	cookie	17	22.7
strawberry ice cream	plan	4	17.3
strawberry ice cream	product	4	17.3
strawberry ice cream	sale	4	17.3
minerals	mineral	39	16.8
lemon tea	tea	40	16.0
lemon tea	lemon	34	15.6
pine juice	pine	43	13.9
N/A	N	50	13.6
mint jelly	mint	58	12.5
mini cake	plan	3	12.0
mini cake	sale	3	12.0
mini cake	product	3	12.0
chocolate cookie	cookie	6	11.9

Figure 7-4 Facet Pairs view showing the Product facet and Noun facet selected

In addition to keywords, you can see the verbs that are highly correlated with a facet. Figure 7-5 shows the result when you select the Product facet and the Verb facet and sort them by correlation.



Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns:

Rows = Product

Columns = Verb

Correlation Amount: Low High

Rows:Product	Columns:Verb	Frequency	Correlation 1 ▼
cookie	sell	4	11.0
cookie	stop	4	11.0
minerals	see	10	10.5
chocolate cookie	dirty	6	9.7
pine juice	juice	10	9.5
fruit jelly	jelly	29	8.2
lemon tea	peel	11	8.0
mint jelly	mint	12	7.8
minerals	loosen	12	7.6
pine juice	pine	6	6.3
lemon tea	separate	6	5.6
cookie	go	6	5.5
minerals	think	10	5.4
N/A	low	23	5.3
chocolate muffin	tear	6	5.2

Figure 7-5 Facet Pairs view showing the Product facet and Verb facet selected

Depending on the facets you select and the data you have, you might discover interesting insights without any customization. Alternatively, you might not discover anything that requires special attention immediately.

7.1.3 Considerations about what you want to discover from the data

Even though you can find useful insight from the data with various views, you might also want other perspectives of your data. At this point, only predefined facets, such as Category, Subcategory and Product, are associated with the XML elements (native fields) found in the source XML file.

What if you want to traverse the current data from a different aspect apart from the predefined facets? You can use a custom dictionary or the custom text analysis rules to look into the data from different aspects.

The next section, through the use of scenarios, explains how you can define a dictionary or custom text analysis rules with the sample data to potentially discover additional insights from your data.

7.2 Content analysis scenarios

To better help you how perform content analysis, this section includes the following scenarios:

- Scenario 1: Using a custom dictionary to discover package-related calls
- Scenario 2: Using custom text analysis rules to discover trouble-related calls
- Scenario 3: Discovering the cause of increasing calls

For the procedures to create the associated custom dictionary and custom text analysis rules, see 7.3, “Configuring the Dictionary Lookup annotator” on page 299, and 7.4, “Configuring the Pattern Matcher annotator” on page 309.

Before you proceed with the scenarios, see Chapter 3, “Understanding content analysis” on page 45, if you have not already done so.

7.2.1 Scenario 1: Using a custom dictionary to discover package-related calls

This scenario explains how to use a custom dictionary to discover package-related calls for the call center. The Facet Pairs view is used for this scenario.

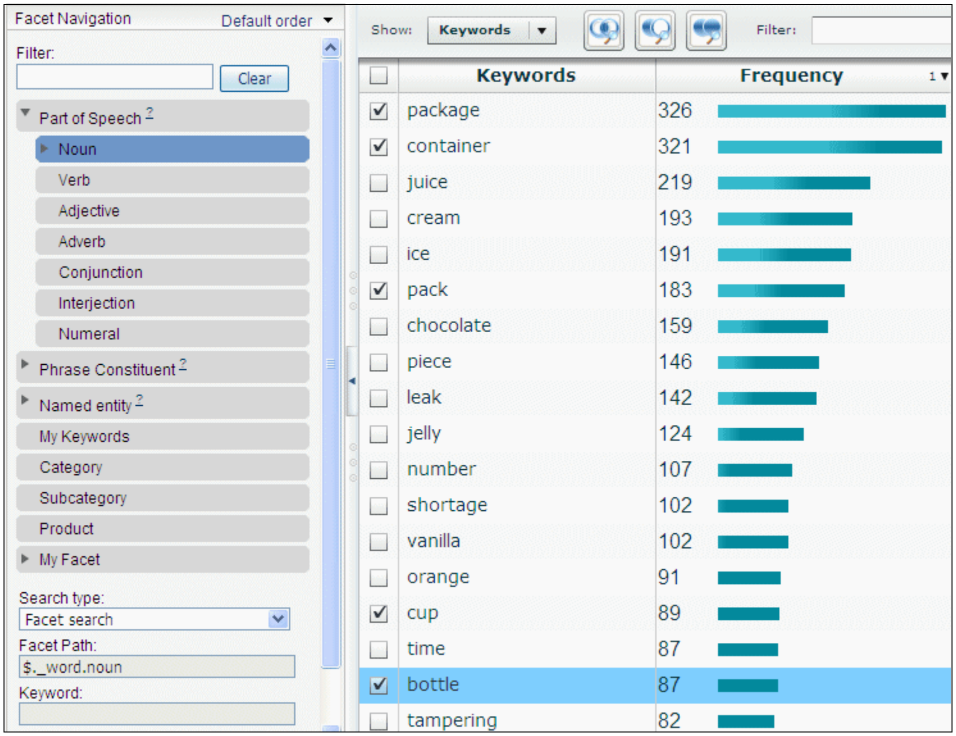
> **Tip for finding candidates for a dictionary entry or rule:** The best resource for finding words and patterns is your collection data. The predefined Part of Speech and Phrase Constituent facets provide good candidates. The Terms of Interest facet also shows a candidate list if it is enabled.

Considering the words to register with the dictionary

First you want to determine the kinds of words that are in the calls that might indicate that they are package-related calls. To determine these words, from the Facet Navigation pane, follow these steps:

1. Expand the **Part of Speech** facet, and select the **Noun** facet.
2. Go to the Facets view. Notice that many words are related to packages. These words include package, container, pack, bottle, and cup, as shown in Figure 7-6.

   Track all the words that you identify that are related to the package calls, and add them to your custom dictionary. We select to track the words **bag**, **bottle**, **cap**, **container**, **cup**, **material**, **pack**, **package**, **shape**, **spoon**, **straw**, and **top**.



Search type: Facet search

Facet Path: $._word.noun

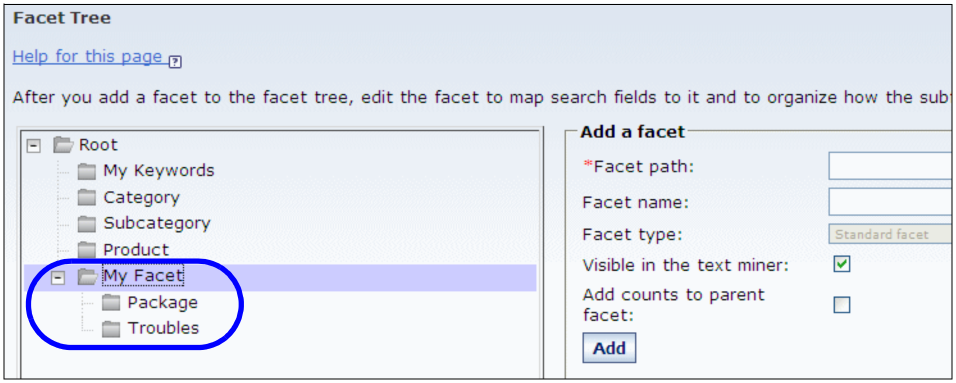
Show: Keywords

☐	Keywords	Frequency
☑	package	326
☑	container	321
☐	juice	219
☐	cream	193
☐	ice	191
☑	pack	183
☐	chocolate	159
☐	piece	146
☐	leak	142
☐	jelly	124
☐	number	107
☐	shortage	102
☐	vanilla	102
☐	orange	91
☑	cup	89
☐	time	87
☑	bottle	87
☐	tampering	82

Figure 7-6 Facets view showing the Noun facet

Updating the facet tree

After you decide the words that you want to add to the custom dictionary, create a facet to associate these words. In this example, we create the My Facet facet to distinguish it from the predefined facets. We create two additional facets under My Facet called “Package” and “Troubles” as shown in Figure 7-7.



Facet Tree

Help for this page

After you add a facet to the facet tree, edit the facet to map search fields to it and to organize how the sub

- Root
  - My Keywords
  - Category
  - Subcategory
  - Product
  - My Facet
    - Package
    - Troubles

Add a facet

*Facet path:

Facet name:

Facet type: Standard facet

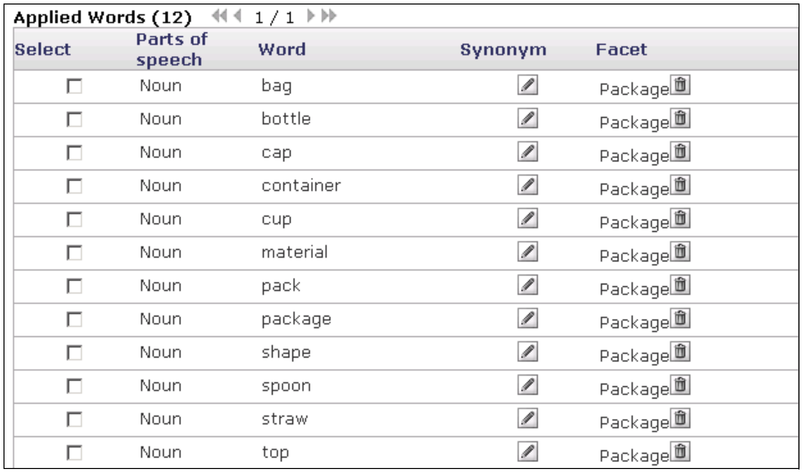
Visible in the text miner: ☑

Add counts to parent facet: □

Figure 7-7 Facet tree after defining the Package facet and Troubles facet

Creating a custom dictionary and associating the words with a facet

Now, you are ready to create a custom dictionary and associate the words that you identified earlier with the new facet. In this scenario, we create a custom dictionary (`package.adic.xml` file). We associate the words that we selected earlier with the Package facet, as shown in Figure 7-8. Now whenever any of the words is displayed in a call, the call is logged with the Package facet.



Applied Words (12) 1 / 1

Select	Parts of speech	Word	Synonym	Facet
□	Noun	bag	✎	Package
□	Noun	bottle	✎	Package
□	Noun	cap	✎	Package
□	Noun	container	✎	Package
□	Noun	cup	✎	Package
□	Noun	material	✎	Package
□	Noun	pack	✎	Package
□	Noun	package	✎	Package
□	Noun	shape	✎	Package
□	Noun	spoon	✎	Package
□	Noun	straw	✎	Package
□	Noun	top	✎	Package

Figure 7-8 Custom dictionary showing associating keywords with the Package facet

To associate special nouns with a particular facet, you define them in the custom dictionary. If you want to associate verbs or phrases with a facet, you use custom text analysis rules as explained in 7.2.2, “Scenario 2: Using custom text analysis rules to discover trouble-related calls” on page 291.

For details about how to create a custom dictionary, see 7.3, “Configuring the Dictionary Lookup annotator” on page 299.

Deploying resources and rebuilding the index

After the dictionary is ready, you must deploy the resources from the administration console. After the resource is deployed successfully, you can see the updated facet tree in the Facet Navigation pane in the text miner application.

The resource deployment does not update the documents that are already indexed. To view the changes in the dictionary, rebuild the index to apply the changes to the data that is indexed already.

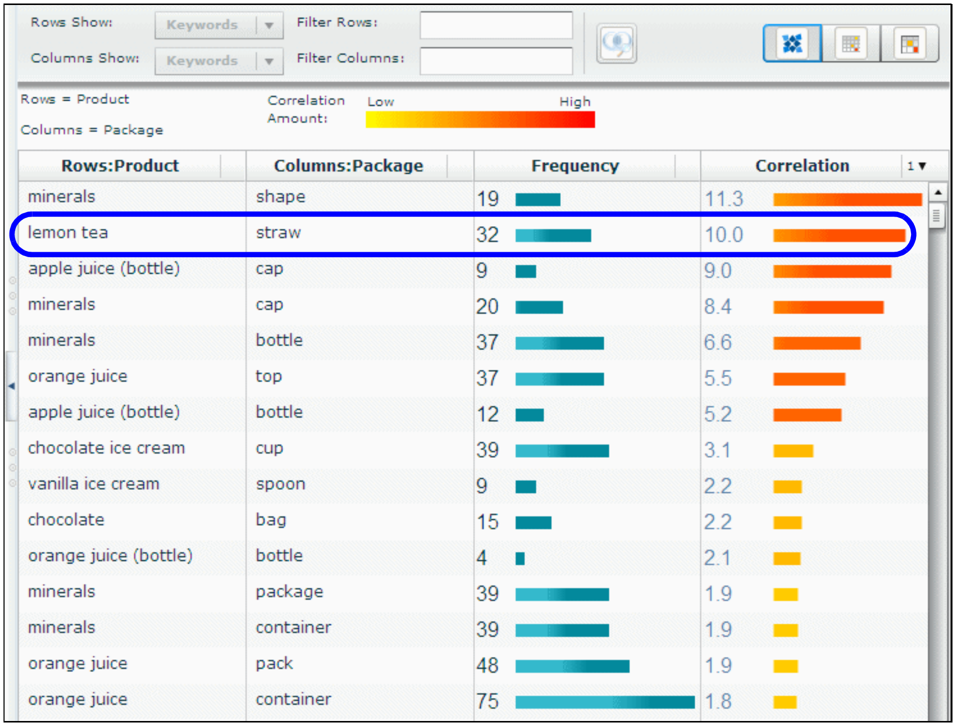
Confirming the result

After you rebuild the index, go back to the text miner application, and confirm the results. If necessary, perform further analysis.

Go to the Facet Pairs view, select the **Product** facet for the row and the **Package** facet for the column. Sort the result by correlation as shown in Figure 7-9 on page 291.

As shown in Figure 7-9 on page 291, a product, such as lemon tea, has a high correlation with the word “straw.” You can look through the documents to see if you can discover any actionable insights. For this scenario, we add the Boolean search condition AND and see the result documents in the Documents view for analysis.

After we review each result document, we notice that several reports indicate that the package of lemon tea did not include a straw with its package or that a straw was separated from the package itself. These reports indicate that the lemon tea product might have a problem with its packaging. This finding is the result from the dictionary definition for this scenario. With a custom dictionary, you can discover the potential defect.



Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns:

Rows = Product

Columns = Package

Correlation Amount: Low High

Rows:Product	Columns:Package	Frequency	Correlation
minerals	shape	19	11.3
lemon tea	straw	32	10.0
apple juice (bottle)	cap	9	9.0
minerals	cap	20	8.4
minerals	bottle	37	6.6
orange juice	top	37	5.5
apple juice (bottle)	bottle	12	5.2
chocolate ice cream	cup	39	3.1
vanilla ice cream	spoon	9	2.2
chocolate	bag	15	2.2
orange juice (bottle)	bottle	4	2.1
minerals	package	39	1.9
minerals	container	39	1.9
orange juice	pack	48	1.9
orange juice	container	75	1.8

Figure 7-9 Facets Pair view showing the Products facet and the Package facet

7.2.2 Scenario 2: Using custom text analysis rules to discover trouble-related calls

This scenario explains how to use custom text analysis rules to discover trouble-related calls for the call center. The Facet Pairs view is used for this scenario.

Considering the patterns to register as rules

As mentioned earlier, the dictionary works with nouns. When you want to find something related to specific verbs or keyword patterns, you use the pattern matching feature of Content Analytics.

To determine the verbs or phrases that you want to consider for pattern matching, expand the **Part Of Speech** (POS) facet, and in the Facet Navigation pane, select the **Verb** facet. Then go to the Facets view.

We notice verbs, such as leak, smell, dirty, loosen, and lower, that indicate some kind of trouble. Therefore, you want to identify the verbs that indicate trouble. For our scenario, we select the verbs leak, smell, dirty, loosen, lower, peel, expire, clump, and detach as trouble words.

In the custom text analysis rule, you can also add nouns as patterns. For our scenario, we select the nouns leak, shortage, contamination, tampering, hole, dirt, prank, clump, hair, thread, empty, bruise, nail, and chunk, which also indicate trouble.

Updating the facet tree

After you decide the verbs and nouns for the custom text analysis rules, create a facet to associate it with the custom text analysis rules. In this scenario, we want to create a Troubles facet to associate pattern matching. Because we already created it earlier, we do not need to update the facet tree at this time. Figure 7-7 on page 289 shows the facet tree.

Creating custom text analysis rules

From the administration console, you can add the custom text analysis rules. However, before adding the custom text analysis rules, you must create a rule file that is written in XML. You can create your own rule file based on keywords or phrases.

> **Facet path:** If you are uncertain about the exact facet path to use as the category attribute in the <mi> element, check the Facet Navigation pane. The Facet Navigation pane shows what the facet path is.

In this scenario, we create the `trouble.pat` rule file (Example 7-2).

Example 7-2 The trouble.pat rule file

```
<?xml version="1.0" encoding="UTF-8"?>
<pattern-list lang="en">
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="leak"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="smell"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="/^dirt/"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="loosen"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
```

```
 <w id="1" lex="lower"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="peel"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="expire"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="clump"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="detach"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="shortage"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="contamination"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="tampering"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="hole"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="prank"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="clump"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="hair"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="thread"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="empty"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="bruise"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
```

```
 <w id="1" lex="nail"/>
</mi>
<mi category="$.myfacet.troubles" value="${1.lex}">
 <w id="1" lex="chunk"/>
</mi>
</pattern-list>
```

Tip to create a rule pattern file: To create a rule pattern file, start with an existing rule pattern file and make a copy of the file. Then edit from the copy and use it in your system. You do not need to create the file from scratch.

For details about how to configure the custom text analysis rules using the user interface, see 7.4, “Configuring the Pattern Matcher annotator” on page 309.

Deploying resources and rebuilding the index

After the dictionary and pattern rules are ready, deploy the resources from the administration console. When the resource is deployed successfully, you can see the updated facet tree in the Facet Navigation pane of the text miner application.

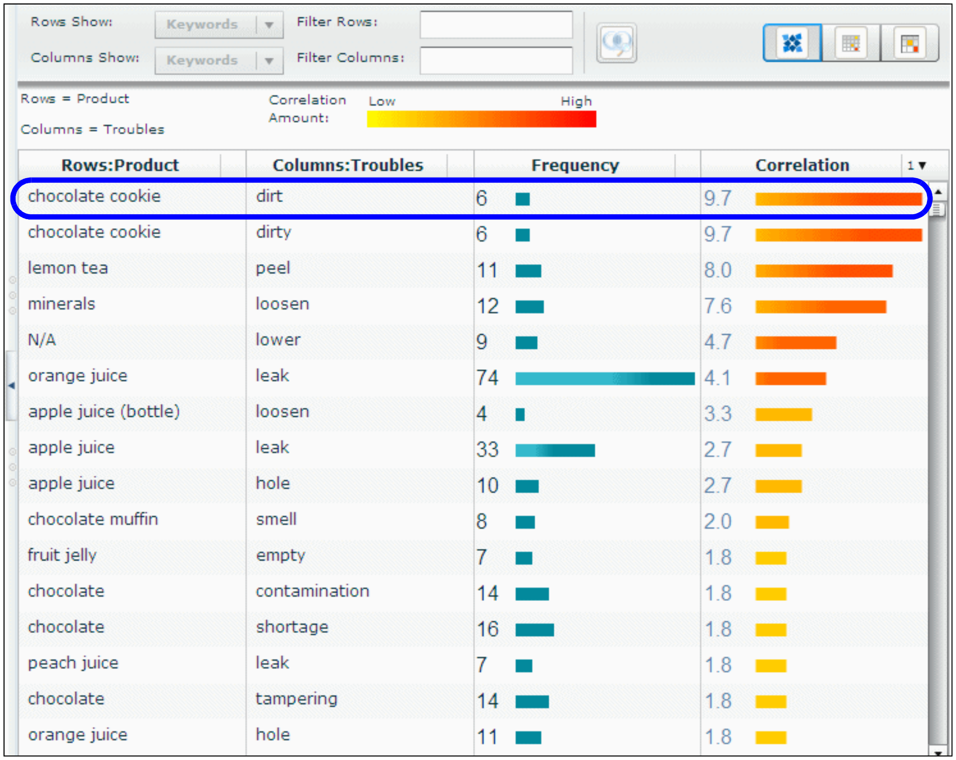
The resource deployment does not update the documents that are already indexed. To reflect the changes in the dictionary, you must rebuild the index to apply the changes to the data that is indexed already.

Confirming the result

After the index is rebuilt, go back to the text miner application, and confirm the results. If necessary, perform further analysis.

Go to the Facet Pairs view again. Select the **Product** facet for row and the **Troubles** facet for column. Sort the result by correlation.

As shown in Figure 7-10, you can see that a product, such as chocolate cookie, has a high correlation with words such as “dirty” or “dirt.”



Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns:

Rows = Product

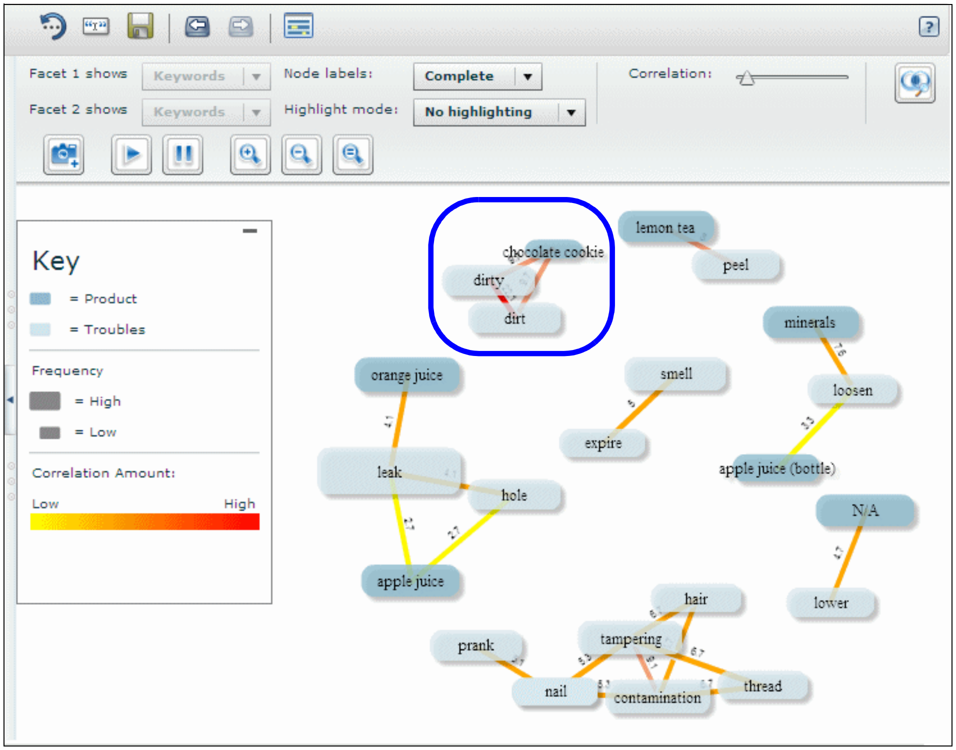
Columns = Troubles

Correlation Amount: Low High

Rows:Product	Columns:Troubles	Frequency	Correlation
chocolate cookie	dirt	6	9.7
chocolate cookie	dirty	6	9.7
lemon tea	peel	11	8.0
minerals	loosen	12	7.6
N/A	lower	9	4.7
orange juice	leak	74	4.1
apple juice (bottle)	loosen	4	3.3
apple juice	leak	33	2.7
apple juice	hole	10	2.7
chocolate muffin	smell	8	2.0
fruit jelly	empty	7	1.8
chocolate	contamination	14	1.8
chocolate	shortage	16	1.8
peach juice	leak	7	1.8
chocolate	tampering	14	1.8
orange juice	hole	11	1.8

Figure 7-10 Facets Pair view showing the Products facet and the Trouble facet

You can also see the relationships between a pair of facets by using the Connections view as shown in Figure 7-11.



Facet 1 shows: Keywords
Facet 2 shows: Keywords
Node labels: Complete
Highlight mode: No highlighting
Correlation:

Figure 7-11 Connections view showing the relationships between the Product facet and the Troubles facet

We again add the condition as a Boolean search AND, and review the result documents in the Documents view for details. After we review each document, we notice that some reports indicate that the container inside was dirty for some reason. With pattern matching, you can discover potential problems with the container.

7.2.3 Scenario 3: Discovering the cause of increasing calls

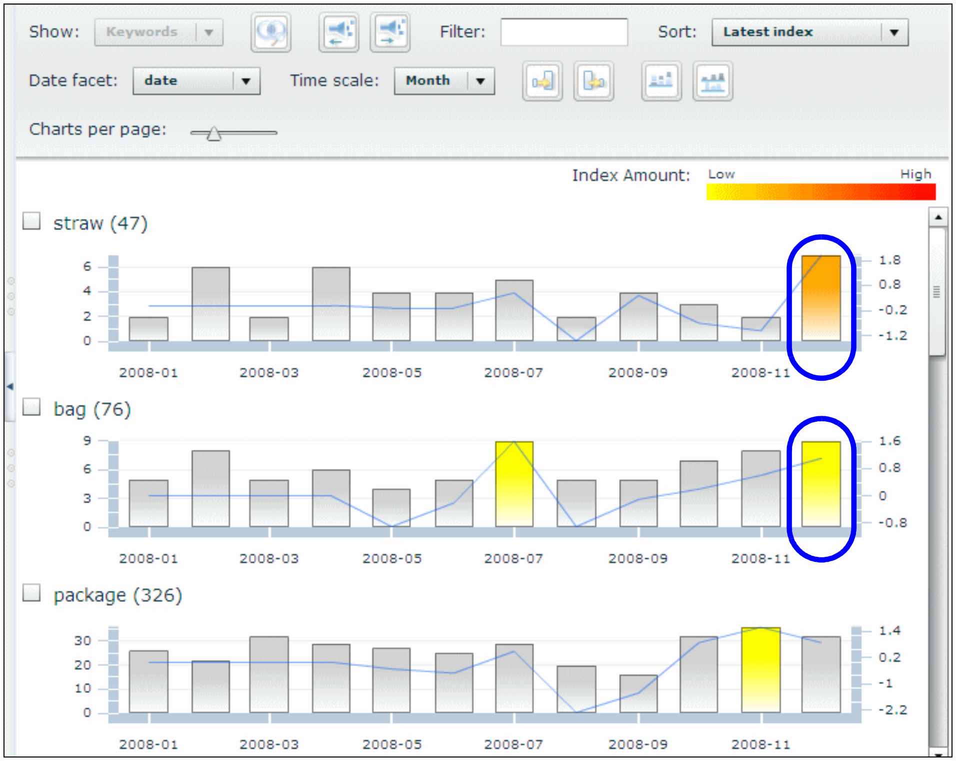
This scenario is a bit different from the first two scenarios. As shown in Figure 7-3 on page 284, when you see the Trends view with the Product facet sorted by the latest index, you see that the calls related to pine juice increase in December 2008. A sharp increase in the Trends view usually indicates something that you must investigate.

In this scenario, we consider investigating the cause of the increasing calls by using the dictionary that we defined previously to see if we can find anything noticeable from the product package aspect.

> **Custom dictionary:** We use the same custom dictionary as defined in 7.2.1, “Scenario 1: Using a custom dictionary to discover package-related calls” on page 287. If you want to analyze the data from a different aspect, you can update your custom dictionary.
>
> In this scenario, we consider the possible cause of the increasing calls for pine juice. For this purpose, package-related calls might be helpful because “package” is commonly used with other products, and there might be a correlation here.

Confirming the result of Scenario 3

In the Trends view with the package facet, we noticed that the calls related to straw and bag increased in December of 2008, as shown in Figure 7-12.



Show: Keywords

Filter:

Sort: Latest index

Date facet: date

Time scale: Month

Charts per page:

Figure 7-12 Trends view showing the Package facet sorted by the latest index

With the custom dictionary defined earlier, we have another analytical aspect of the package that can be related to pine juice and other products if those items are commonly used as pine juice. We discover the potential package-related problem with the Trends view, with the help of pattern matching and custom dictionary.

7.2.4 Conclusion

As you can see by the scenarios in this section, you can use the text miner application to determine the symptom or the trend. You can also use the text miner application to determine what to analyze, select the candidates for the custom dictionary, or define the custom text analysis rules to discover the insight from various aspects. You can also use the text miner application to discover noticeable events, detect anomalies at the beginning when the problem occurs, or predict future trends.

As shown in this section, when you perform analysis with the text miner application, you typically perform the following steps:

1. Analyze the data, and review the results with the text miner application.
2. Consider which words you want to analyze. That is, select the candidate words for the custom dictionary or the custom text analysis rules.
3. Define a facet tree so that you can see the data from the specific aspect.
4. Define a custom dictionary or the custom text analysis rules, and associate them with the facet.
5. Deploy the resources, and rebuild the index.
6. Confirm the result with the text miner application. See if you can discover insights from a defined custom dictionary or custom text analysis rules.

   If not, consider iterating the process until you can discover what you want with different dictionary keywords or different custom text analysis rules.

The analysis is an iterative process. You need to employ a trial-and-error approach until you gain interesting insights that help your business.

Remember that the text miner application does not provide what you need to focus on automatically. You must be conscientious of how you want to create the custom dictionary or the custom text analysis rules with a given data set based on what you want to analyze. For more details about this concept, read Chapter 3, “Understanding content analysis” on page 45.

The rest of this chapter provides details for creating and editing a custom dictionary and custom text analysis rules by using the Dictionary Lookup annotator and Pattern Matcher annotator.

7.3 Configuring the Dictionary Lookup annotator

The Dictionary Lookup annotator matches words and synonyms from dictionaries with words in your text. The annotator also associates the keywords with user-defined facets.

Keywords are a critical element in text analysis. When you know particular terms in a specific domain, for example, product names, they are useful for extracting documents that belong to a specific domain. Keywords can be grouped by concept type and then used to identify documents with interesting combinations of these concepts.

The Dictionary Lookup annotator finds user-defined keywords in a document and associates the words with user-specified facets. The Dictionary Lookup annotator is a simple but powerful way to identify particular keywords.

Dictionary Lookup annotator is enabled by default. If you use the Dictionary Lookup annotator, you must also enable the Pattern Matcher annotator.

Dictionary Editor: The Dictionary Editor supports nouns only. You cannot use the Dictionary Editor to add other parts of speech such as verbs, adjectives, or adverbs. To capture other parts of speech, such as verbs and adjectives, use the Pattern Matcher annotator.

Noun identification: A document has zero to many fields. Each field has some attributes such as analyzable. For noun identification, text analytics is applied to all analyzable fields and content. XML documents are processed as content and thus can also be analyzed.

7.3.1 When to use the Dictionary Lookup annotator

Consider using the Dictionary Lookup annotator in the following cases:

- You want to see particular noun terms as a facet in the text mining application.
- You want to add new nouns to enhance the linguistic analysis process.
- You already use the dictionary in IBM Content Analyzer, the predecessor product of Content Analytics.

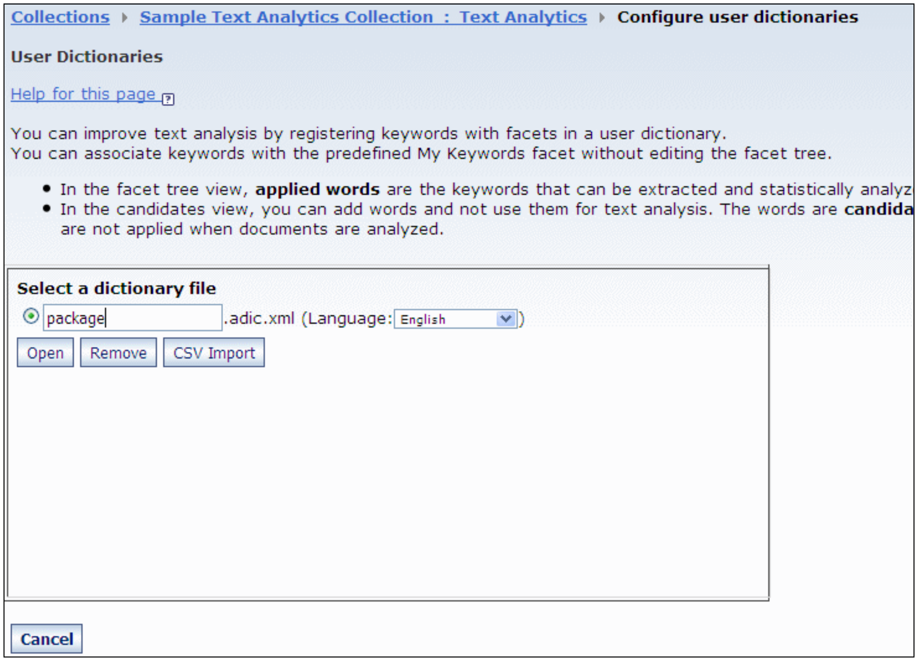
See 7.2.1, “Scenario 1: Using a custom dictionary to discover package-related calls” on page 287, for the reason why you might want to use the Dictionary Lookup annotator to create a custom dictionary.

7.3.2 Configuring custom user dictionaries

To create a custom user dictionary and to add, edit, and delete keywords and their synonyms, you can use the administration console. Using the scenario described in 7.2.1, “Scenario 1: Using a custom dictionary to discover package-related calls” on page 287, we show how to add the nouns **bag**, **bottle**, **cap**, **container**, **cup**, **material**, **pack**, **package**, **shape**, **spoon**, **straw**, and **top** that we selected from that scenario.

To add the nouns to your custom dictionary, follow these steps:

1. From the administration console, click **Parse and Index** from the collection to which you want to add the custom user dictionary.
2. Select **Text Analytics** → **Edit** → **Configure user dictionaries**.
3. In the Configure user dictionaries panel (Figure 7-13), complete these steps:
   a. Enter the custom dictionary name. We type `package`.
   b. For Language, select the language that the dictionary will be applied to. The drop-down list shows the languages that you selected when creating a collection. We select **English**.
   c. Click **Open**.



Collections ▸ Sample Text Analytics Collection : Text Analytics ▸ **Configure user dictionaries**

User Dictionaries

Help for this page

You can improve text analysis by registering keywords with facets in a user dictionary.
You can associate keywords with the predefined My Keywords facet without editing the facet tree.

- In the facet tree view, **applied words** are the keywords that can be extracted and statistically analyz
- In the candidates view, you can add words and not use them for text analysis. The words are **candida** are not applied when documents are analyzed.

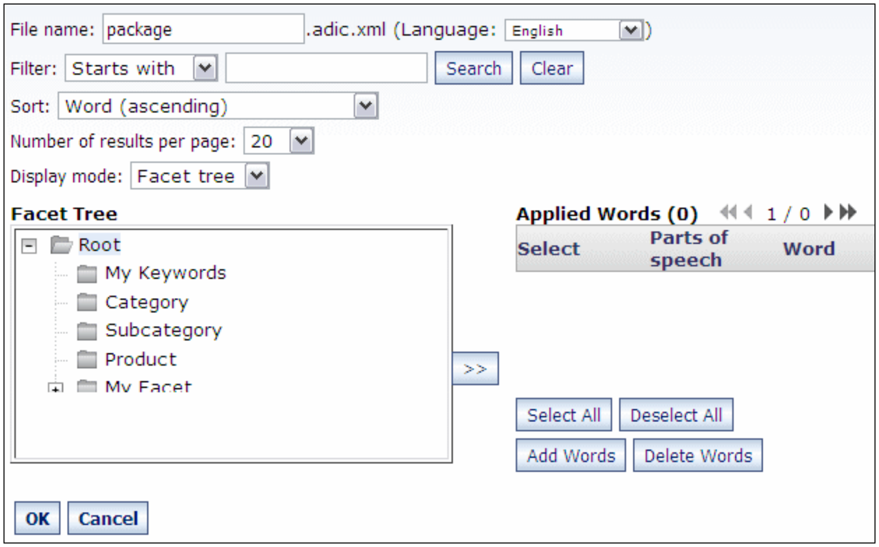
Select a dictionary file

● package .adic.xml (Language: English)

Figure 7-13 User Dictionaries configuration window

4. In the Dictionary Editor (Figure 7-14), create new keywords:

   a. Click **Add Words**.



File name: package .adic.xml (Language: English)

Filter: Starts with

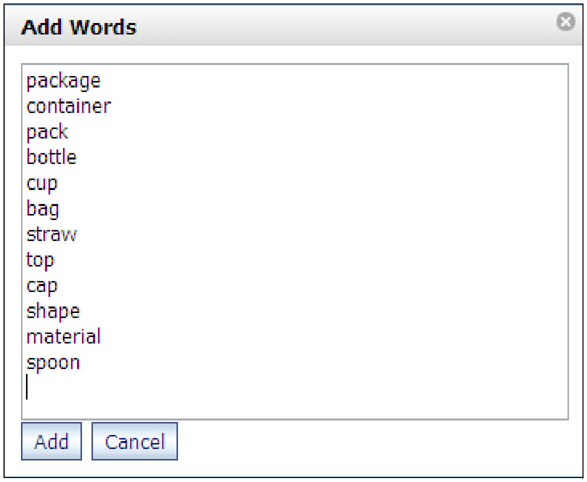
Sort: Word (ascending)

Number of results per page: 20

Display mode: Facet tree

Figure 7-14 Dictionary Editor

   b. In the Add Words window (Figure 7-15), enter keyword strings. You can add multiple keywords at a time. The format is one keyword per line. We enter `package`, `container`, and other keywords. Click **Add**.

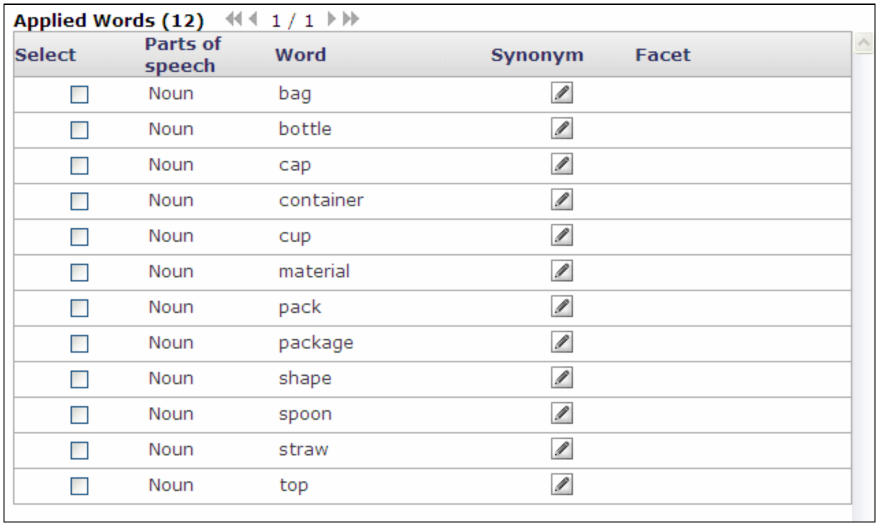


Add Words

```
package
container
pack
bottle
cup
bag
straw
top
cap
shape
material
spoon
```

Figure 7-15 Adding new keywords in the dictionary

The keyword editor reflects the new keywords as shown in Figure 7-16.



Applied Words (12) 1 / 1

Select	Parts of speech	Word	Synonym	Facet
□	Noun	bag	✎	
□	Noun	bottle	✎	
□	Noun	cap	✎	
□	Noun	container	✎	
□	Noun	cup	✎	
□	Noun	material	✎	
□	Noun	pack	✎	
□	Noun	package	✎	
□	Noun	shape	✎	
□	Noun	spoon	✎	
□	Noun	straw	✎	
□	Noun	top	✎	

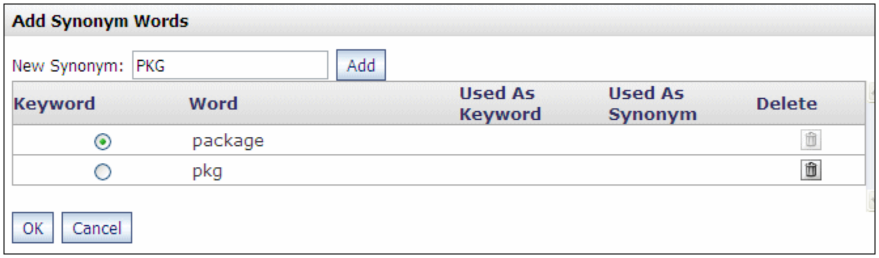
Figure 7-16 Dictionary Editor showing the new keywords

In many cases, keywords come with alias names and abbreviations. Keywords can also have variant forms (inflected forms). You can add multiple synonyms that have an identical meaning for a defined keyword. The Dictionary Lookup annotator captures synonyms and treats them as a single keyword.

5. To create a synonym for a keyword, from the Applied Words list (Figure 7-16), follow these steps:

   a. Click the **Edit Synonym** icon for a given keyword.
   b. Enter a new synonym in the top text box. Click **Add**.
   c. Repeat these steps until you add all synonyms for the given keyword.

   For example, we add the lowercase abbreviation `pkg` and the uppercase abbreviation `PKG` to the keyword `package` as shown in Figure 7-17.



Add Synonym Words

New Synonym: PKG Add

Keyword	Word	Used As Keyword	Used As Synonym	Delete
●	package			🗑
○	pkg			🗑

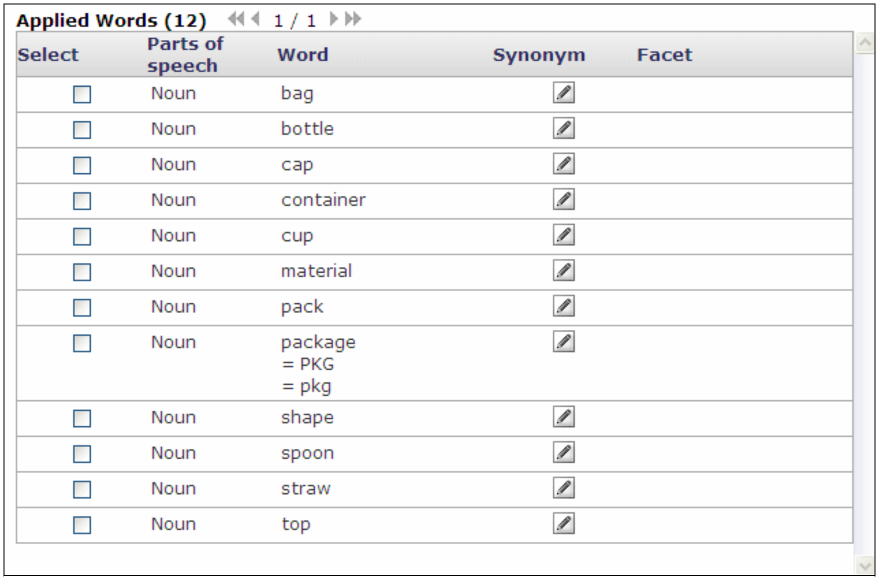
OK Cancel

Figure 7-17 Adding synonyms for a keyword in the dictionary

In Figure 7-17 on page 302, the first entry, `package`, is selected as the keyword, which means that `package` is the normal form of these synonyms. If a word is already added as a normal form of the other term, then Used As Keyword is selected. If a word is already added as a synonym for the other term, Used As Synonym is selected. These two features are warning signs for conflicts with other entries.

d. Click **OK**.

Dictionary Editor now reflects your synonyms (Figure 7-18).



Applied Words (12) 1 / 1

Select	Parts of speech	Word	Synonym	Facet
□	Noun	bag	✎	
□	Noun	bottle	✎	
□	Noun	cap	✎	
□	Noun	container	✎	
□	Noun	cup	✎	
□	Noun	material	✎	
□	Noun	pack	✎	
□	Noun	package = PKG = pkg	✎	
□	Noun	shape	✎	
□	Noun	spoon	✎	
□	Noun	straw	✎	
□	Noun	top	✎	

Figure 7-18   Dictionary Editor showing synonyms (package)

Synonyms: Synonyms are not displayed as discrete words with their associated facet in the text miner application. Only the keyword (not its synonyms) is displayed (in this example, `package`) in the Facets view.

6. Associate the defined keywords to a facet. If you have not created a facet, follow the steps in “Creating facets and mapping search fields to facets” on page 106.

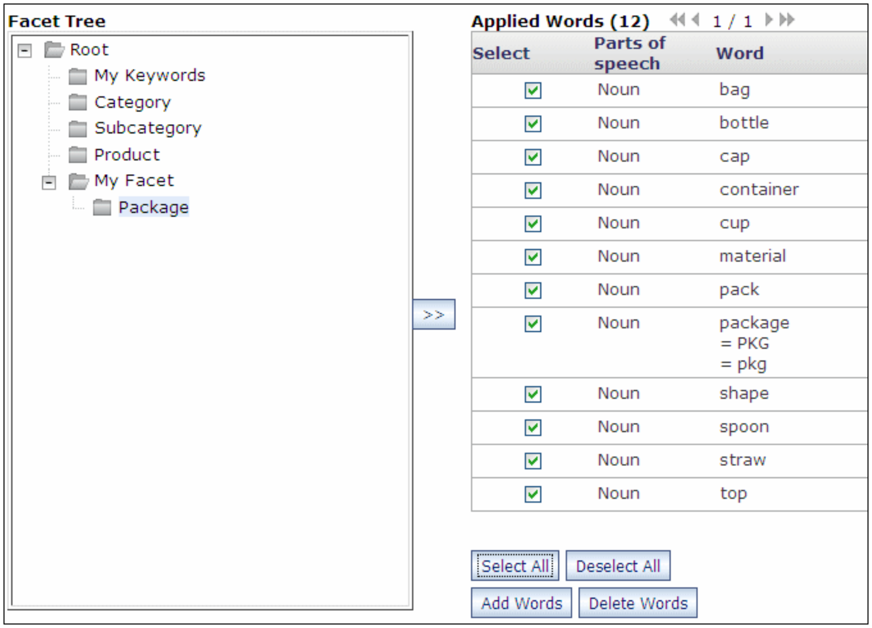
   To associate keywords to a facet, follow these steps:

   a. In the Facet Tree view, select a facet.

   b. In the Applied Words table, select the keywords that are associated with the selected facet.

c. Click the **Assign** button (>>).

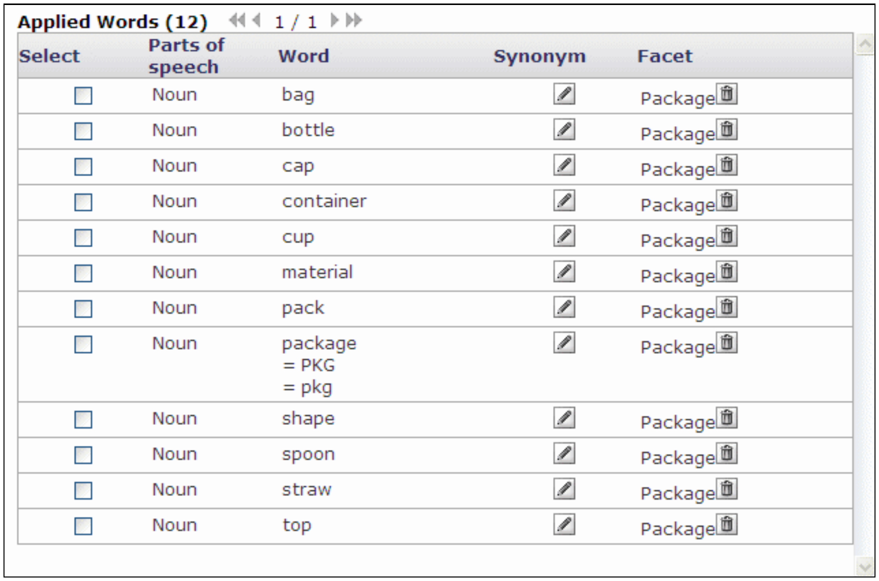
In this example, we assign all keywords to the Package facet under My Facet as shown in Figure 7-19.



Select	Parts of speech	Word
☑	Noun	bag
☑	Noun	bottle
☑	Noun	cap
☑	Noun	container
☑	Noun	cup
☑	Noun	material
☑	Noun	pack
☑	Noun	package = PKG = pkg
☑	Noun	shape
☑	Noun	spoon
☑	Noun	straw
☑	Noun	top

Figure 7-19 Associating keywords to a facet

Dictionary Editor now reflects the facet that you associated with the keywords (Figure 7-20).



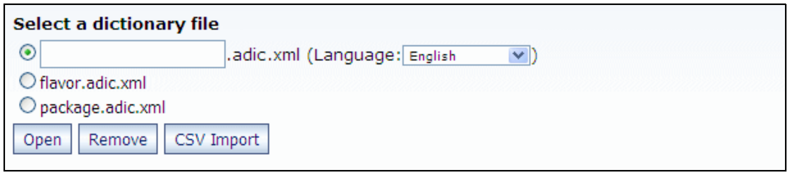
Applied Words (12) 1 / 1

Select	Parts of speech	Word	Synonym	Facet
□	Noun	bag	✎	Package
□	Noun	bottle	✎	Package
□	Noun	cap	✎	Package
□	Noun	container	✎	Package
□	Noun	cup	✎	Package
□	Noun	material	✎	Package
□	Noun	pack	✎	Package
□	Noun	package = PKG = pkg	✎	Package
□	Noun	shape	✎	Package
□	Noun	spoon	✎	Package
□	Noun	straw	✎	Package
□	Noun	top	✎	Package

Figure 7-20 Dictionary Editor showing the facets

d. Click **OK** to save the dictionary.

7. If you have other word lists that you identified from another perspective, add multiple dictionaries for a collection. Repeat step 3 on page 300 through step 6. In Figure 7-21, we created another dictionary to capture keywords that are related to "flavor."



Select a dictionary file

● .adic.xml (Language: English)

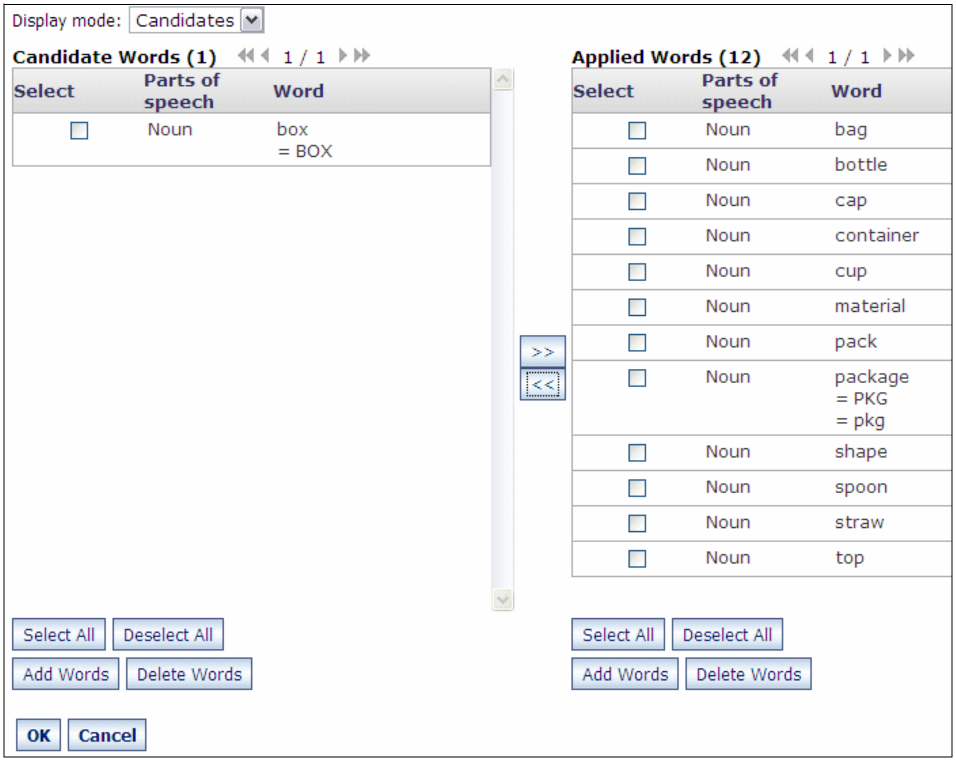
○ flavor.adic.xml

○ package.adic.xml

Figure 7-21 Selecting the flavor dictionary

Conflicts: Dictionary Editor only checks conflicts within the same dictionary. It is important to store keywords that belong to the same facet in a single dictionary.

Applied words are the keywords that are used for text analysis. If the words fall into the candidate words list, these words are not used for text analysis. Moving words to the Candidate Words list is useful when evaluating potential words to include in your dictionaries. Figure 7-22 shows the Candidates mode. In this example, we move the keyword “box” from the Applied Words list to the Candidate Words list. This way, the noun “box” will not be used in future analysis. In moving a keyword from the Applied Words box, notice that the associated synonyms also move with the selected word.



Display mode: Candidates

Candidate Words (1) 1 / 1

Select	Parts of speech	Word
□	Noun	box = BOX

Applied Words (12) 1 / 1

Select	Parts of speech	Word
□	Noun	bag
□	Noun	bottle
□	Noun	cap
□	Noun	container
□	Noun	cup
□	Noun	material
□	Noun	pack
□	Noun	package = PKG = pkg
□	Noun	shape
□	Noun	spoon
□	Noun	straw
□	Noun	top

Figure 7-22 Candidate mode of Dictionary Editor

In addition to adding keywords and their synonyms using the Dictionary Editor, you can also import a comma-separated values (CSV) file that lists words that you want to add in a dictionary. To import a CSV file, follow these steps:

1. From the Configure user dictionaries panel (Figure 7-13 on page 300), click **CSV Import**.
2. In the CSV Import window (Figure 7-23 on page 307), complete the following steps:
   a. Specify an Adic file name. You can select an existing adic file to update or create a dictionary.

b. For Language, specify the language that the dictionary will apply when creating a dictionary.

c. Select a CSV file. The first column in each row is a keyword. The rest of the columns are treated as synonyms. Example 7-3 shows the CSV file that contains two keywords (package and container). `PKG` and `pkg` are synonyms of the keyword “package.”

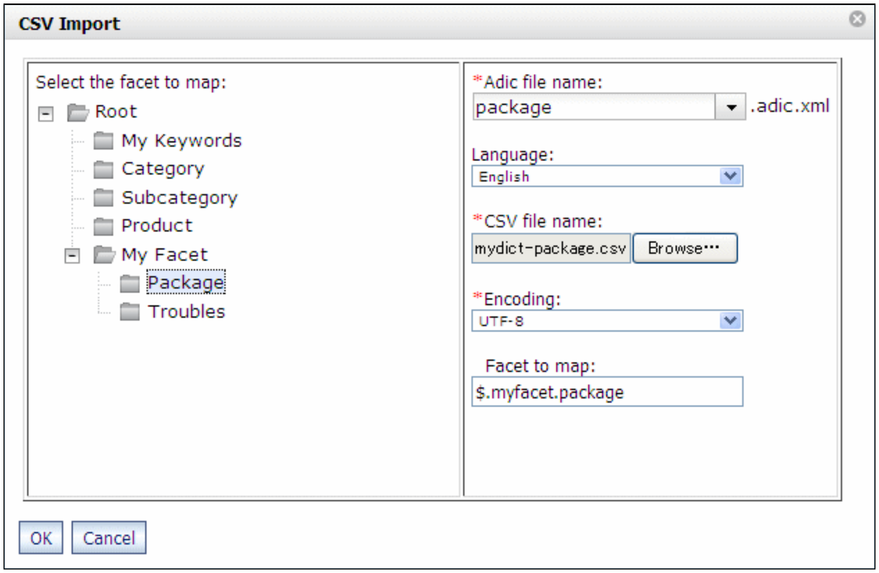
Example 7-3 CSV file that contains two keywords

```
package,PKG,pkg
container
```

d. Select an appropriate encoding for the CSV file.

e. Select a facet to map if necessary.

f. Click **OK**.



CSV Import

*Adic file name: package .adic.xml

Language: English

*CSV file name: mydict-package.csv

*Encoding: UTF-8

Facet to map: $.myfacet.package

Figure 7-23 Importing a CSV file

After importing the CSV file, you can add, edit, and delete keywords and their synonyms using the Dictionary Editor.

For more information about Dictionary Editor, see the “Configuring user dictionaries” topic in the IBM Content Analytics Information Center at the following address:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/topic/com.ibm.discovery.es.ad.doc/iiysatauserdict.htm

After you create a user dictionary, you must deploy the text analytics resources. You must also rebuild an existing index to update the results in the text miner application.

7.3.3 Migrating the Content Analyzer dictionaries

If you use IBM Content Analyzer, which is the predecessor to Content Analytics, and have Content Analyzer user dictionaries, you can use them in Content Analytics.

> **Terminology note:** Content Analytics uses different terminology than Content Analyzer. Content Analyzer uses the term “category,” which is called a “facet” in Content Analytics. See 1.2.2, “Product changes” on page 4, for more changes.

To migrate the Content Analyzer dictionary to the Content Analytics dictionary, follow these steps:

1. Copy the `adic.xml` files to the `ES_NODE_ROOT/master_config/collectionID.indexservice/resource/adic/` directory.
2. Add new facets if necessary.
3. Edit the `adic.xml` files:
   a. Add the **`lang`** attribute to the top-level element dictionary.
   b. Change the facet paths that are represented in the **`cat`** attribute. In Content Analytics, the facet path must start with the dollar sign symbol (`$`).

   Example 7-4 shows the `adic.xml` file with the changes.

   Example 7-4 The adic.xml file in Content Analytics with changes highlighted in bold

   ```
   <dictionary lang="en">
       <entry id="1" cat="$.mykeyword" .../>
   </dictionary>
   ```

4. Deploy the text analytics resources.

7.3.4 Validation and maintenance

The easiest way to confirm that the words in your dictionary match the words in your textual content is to use the Facets view in the text miner application. When you add keywords and associate them with a particular facet, you can see the added keywords in the Facets view if your collection contains documents with these keywords.

You can easily add, edit, and remove keywords by using the Dictionary Editor. Remember to deploy resources and rebuild the index when you update the dictionary so that the latest dictionary is reflected in the text miner application. The rebuild task might take a long time depending on the size and number of documents in your text analytics collection.

> **Tip for building a dictionary:** When building your dictionary, build it iteratively with a small subset of your content. You can start from a small document set and check the results to make sure that they are what you expected. You can also find other keywords to include in your dictionary during this process. Also enhance your dictionary iteratively. When you think you are done with the iteration process, you can then apply your dictionary to the entire collection.

7.4 Configuring the Pattern Matcher annotator

The Pattern Matcher annotator identifies patterns in your text by using the rules that you defined. The annotator also associates the patterns with user-defined facets.

Using keywords for text analysis is a simple and powerful way to identify documents that contain particular keywords and their synonyms. However, using individual keywords alone is not enough to discover and normalize concepts and ideas. For example, the keyword “milk” can help us easily identify all documents that contain the word milk. However, matching single words is not enough to extract the true context of the documents in many other cases. For example, consider the following sentence:

```
The product is not broken.
```

When you try a traditional search with the keywords “product” and “broken,” any document that contains this sentence is returned in the search results. However, this sentence does not indicate a product problem. We want to distinguish concepts that a document presents, such as broken or not broken. In this case, the following rules can help to decode the actual information in a document:

```
[product name] + [be] + [negative term] = product problem
[product name] + [be] + [not] + [positive term] = product problem
```

The first rule states any product name that is followed by a `be` type of verb and a negative term is considered a product problem. The second rule states that any product name that is followed by a `be` type of verb, a `not` type of word, and a positive term is considered a product problem. This product problem is then considered as a keyword phrase that can be associated with a facet and can be used for analysis. In this scenario, documents that contain any of the following sentences are displayed as having a product problem in the analysis:

```
ProductA is broken.
ProductBs are not working.
```

The Pattern Matcher annotator recognizes the sequences of words that are defined in the rules and associates them with specified facets. With the Pattern Matcher annotator, you can add custom rules.

The Pattern Matcher annotator is enabled by default. It produces the Part of Speech and Phrase Constituent facets, which are predefined by default.

Disabling Pattern Matcher: If you disable the Pattern Matcher annotator, the predefined Part of Speech and Phrase Constituent facets do not show any results.

7.4.1 When to use the Pattern Matcher annotator

For reasons why you might want to use the Pattern Matcher annotator, see 7.2.2, “Scenario 2: Using custom text analysis rules to discover trouble-related calls” on page 291.

In general, you use the Pattern Matcher annotator in the following cases:

- You want to extract sequences of words (single word and multiple words).
- You want to capture patterns that are constructed by multiple words.
- You already have rules defined and used by the predecessor of the Content Analytics product, Content Analyzer.

An alternative approach to use Pattern Matcher annotator is using LanguageWare Resource Workbench. See 11.2.1, “LanguageWare Resource Workbench” on page 454, for more information.

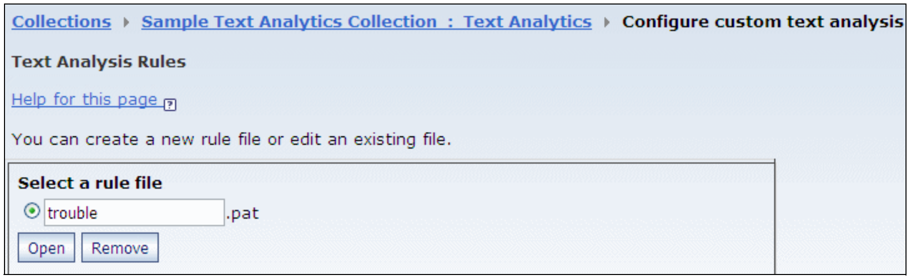
7.4.2 Configuring custom text analysis rules

To construct rules for Pattern Matcher annotator, a certain degree of linguistic knowledge is required.

To create a custom rule file, and edit text analysis rules, you can use the administration console. Using the scenario in 7.2.2, “Scenario 2: Using custom text analysis rules to discover trouble-related calls” on page 291, we add the rules as shown in Example 7-2 on page 292 to extract terms that are possible signs of troubles.

To create the rules, follow these steps:

1. From the administration console, click **Parse and Index** of the collection that you want to add a custom rule file.
2. Select **Text Analytics** → **Edit** → **Configure custom text analysis rules**.
3. In the Text Analysis Rules window (Figure 7-24), enter the custom rule file name and click **Open**.


Collections ▸ Sample Text Analytics Collection : Text Analytics ▸ Configure custom text analysis

Text Analysis Rules

Help for this page

You can create a new rule file or edit an existing file.

Select a rule file

● trouble .pat

Figure 7-24 Text Analysis Rules window showing a rule configuration

4. Add the custom rules. Example 7-5 shows the rule structure in the `.pat` format.

Example 7-5 Pattern rule syntax

```
<pattern-list lang="en">
  <mi category="$.myfacet.favorable" value="able to ${4.lex}">
   <w id="1" lex="be" str="!/^((was)|(WAS))$/"/>
   <w id="2" lex="able"/>
   <w id="3" lex="to"/>
   <w id="4" pos="verb"/>
  </mi>
</pattern-list>
```

To create the XML to define the rules, follow these steps:

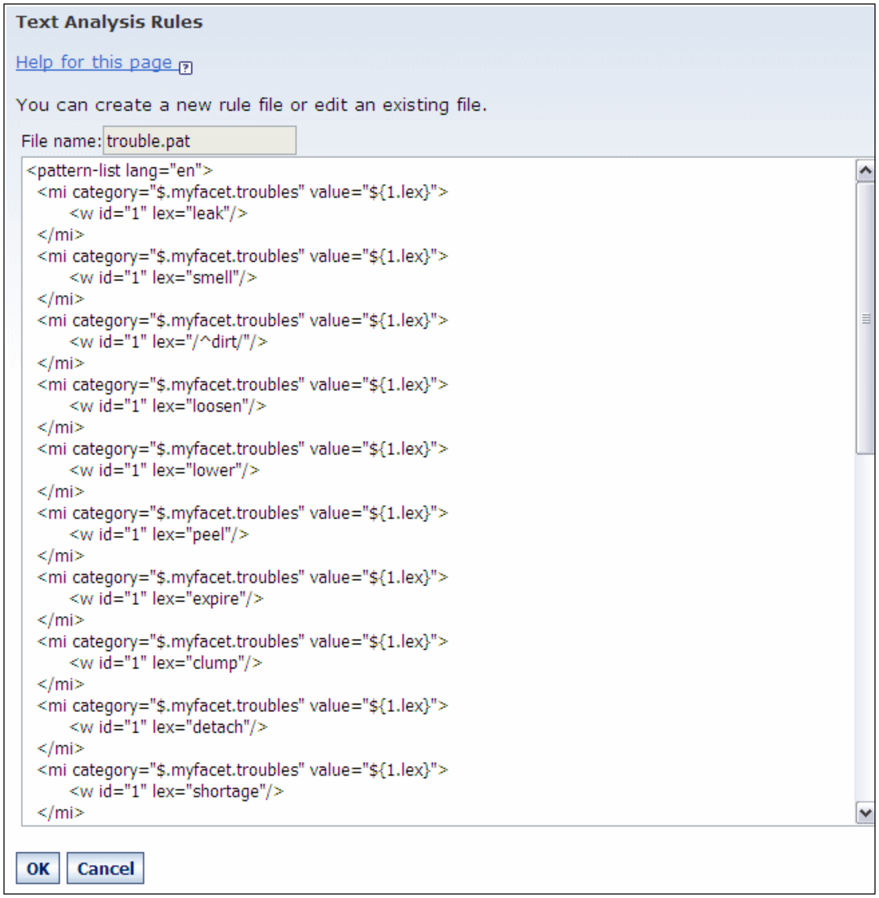
a. Add the top-level element pattern list, which must specify the target language as the `lang` attribute.

b. Add a `<mi>` element that represents a rule. You need to specify the facet path as the `category` attribute. The facet path must start with `$`. The dot (`.`) is the path separator. For example, for the word `favorable` under `myfacet`, use `$.myfacet.favorable`. Also give the `value` attribute that is shown in the Text Mining Application as the keyword.

   In a `<mi>` element, the actual pattern consists of one or more `<w>` elements that represent tokens. The `<w>` element must have a token ID. You can add several constraints by using the `str`, `lex`, `pos`, `ftrs`, `category`, and `guard` attributes. With the Pattern Matcher annotator, you can use regular expression matching in the constraints.

For more information about the rule development, see 7.4.4, “Designing the custom text analysis rules” on page 313.

Figure 7-25 shows the added custom rules in the administration console.



File name: trouble.pat

```
<pattern-list lang="en">
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="leak"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="smell"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="/^dirt/"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="loosen"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="lower"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="peel"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="expire"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="clump"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="detach"/>
  </mi>
  <mi category="$.myfacet.troubles" value="${1.lex}">
       <w id="1" lex="shortage"/>
  </mi>
```

Figure 7-25 Text Analytics Rules Editor

c. Click **OK** to save the rule file.

You can add multiple rule files for a collection.

After creating the rule file, deploy the text analytics resources. In addition, rebuild an existing index to update the results in the text miner application.

Hints: Copy the sample pattern file, and edit it by using an appropriate application that can assist in XML editing. Writing analytic rules from scratch can cause problems even though the XML syntax is simple. The sample pattern files for Voice-of-Customer (VOC) analysis are in the `ES_INSTALL_ROOT/samples/voc/pattern/` directory.

7.4.3 Migrating the Content Analyzer rules

If you use Content Analyzer and have the rules, you can use them in Content Analytics. To migrate Content Analyzer rules to be usable by Content Analytics, follow these steps:

1. Copy the `.pat` files to the `ES_NODE_ROOT/master_config/collectionID.indexservice/resource/pattern/` directory.
2. Add the new facets if necessary.
3. Edit the `.pat` file:
   a. Add the `lang` attribute to the top-level element.
   b. Change the facet paths that are represented in the `category` attribute. In Content Analytics, the facet path must start with `$`.

   Example 7-6 shows the `.pat` file with the changes.

   Example 7-6 The .pat file in Content Analytics with changes highlighted in bold

```
<pattern-list lang="en">
  <mi category="$.pattern.general" value="${1.lex}">
    <w id="1" ... />
  </mi>
</pattern-list>
```

4. Deploy the text analytics resources.

7.4.4 Designing the custom text analysis rules

By defining your own text analysis rules, you can extract concepts that are expressed in natural language.

This section focuses on designing custom text analysis rules. Here we do not explain the details of the text analysis rule syntax. Rather, we pick up several situations in our analysis of the voice of customer and provide some perspectives about text analysis rule development. For more information about the custom analysis rule syntax, see the “Custom rule files for text analytics collections” topic in the IBM Content Analytics Information Center at the following address:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/topic/com.ibm.discovery.es.ta.doc/iiysatextanalrules.htm

You can also see the practical pattern files for voice of customer analysis that are in the `ES_INSTALL_ROOT/samples/voc/pattern/` directory.

Sample rule: Capturing question-related information

In VOC data, you often encounter questions posed by the customer. For example, you might find that a customer says “I need information about ...”. This phrase seems to be a common request. Therefore, you want to capture these types of questions where customers have asked for information associated with a particular subject. You want to create a text analysis rule that gathers these kinds of phrases. You start by creating a custom text analysis rule to capture this simple expression as shown in Example 7-7.

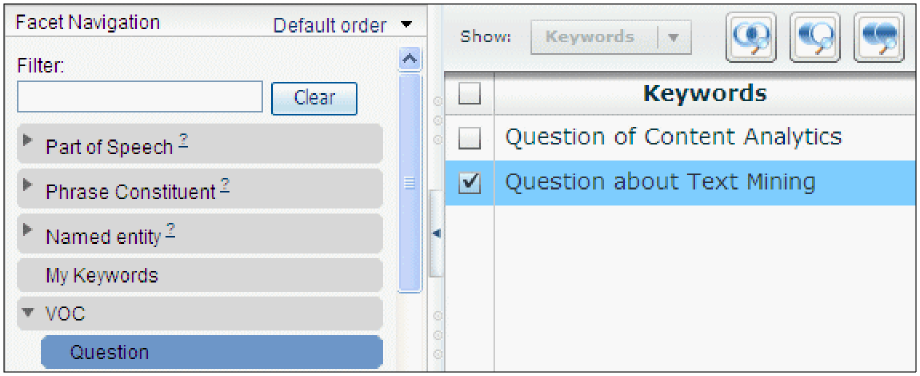
Example 7-7 Simple rule to capture the expression “I need information”

```
<mi category="$.voc.question" value="Question ${4.str} ${5.str}
${6.str}">
  <w id="1" str="I"/>
  <w id="2" str="need"/>
  <w id="3" str="information"/>
  <w id="4"/>
  <w id="5"/>
  <w id="6"/>
</mi>
```

Each <w> element defines a token to be extracted and analyzed. Tokens #1, #2, and #3 use a simple string constraint that matches an actual word. Token #4, #5, and #6 do not have any constraint, meaning they can be any token in a sentence.

In the <mi> element, you must specify the facet path as the category attribute and the value attribute that is showing in the application as the keyword. The value attribute is a template of the output keyword.

After you save this rule within the Text Analytics Rules Editor, you must apply it and verify results. In this case, assume that the pattern matcher annotator examines the sentence “I need information about Text Mining.” You can see that the annotator captures this sentence and assigns it to the VOC/Question facet with a keyword of “Question about Text Mining” as shown in Figure 7-26 on page 315.



□	Keywords
□	Question of Content Analytics
☑	Question about Text Mining

Figure 7-26 Facets view showing results of the sample rule

You also notice another generic expression that asks a question directly. To see the actual sentences, you can attempt a simple search with the keyword “question” and find several sentences in your collection:

- I got a question regarding ...
- I have a question about ...
- I have several questions on ...

To capture this type of expression, we add a rule as shown in Example 7-8.

Example 7-8 Rule to capture question-related information

```
<mi category="$.voc.question" value="Question ${5.lex} ${6.lex}
${7.lex}">
  <w id="1" str="I"/>
  <w id="2" lex="/(^have$)|(^get$)/"/>
  <w id="3"/>
  <w id="4" lex="question"/>
  <w id="5"/>
  <w id="6"/>
  <w id="7"/>
</mi>
```

The rule starts with a simple string constraint that matches “I.” Token #2 uses the lex attribute that stands for the lemma (linguistic normalized form of the word). With this attribute, you can write a constraint without listing all possible inflected forms of a word. For example, `lex="get"` matches “get,” “gets,” “getting,” “got,” and “gotten” that have the normalized form “get”. The lex value of token #2 is using regular expression matching. All rules with the forward slash (/) operator are evaluated by using `java.util.regex` classes. This lex value uses the pipe (|),

^, $, and () operators, so that it matches “have”, “had”, “get”, and “got”. It does not match “haven” and “forget” for example.

> **The / operator:** The / operator itself is not a part of the Java regular expression syntax. It is used only to invoke regular expression processing on a rule.

Token #3 does not have any constraints. This token can be anything in a sentence. It is a place holder for any article or numeral such as “a,” “the,” “one,” or “some”. Token #4 also uses a lex constraint to ignore a difference between a singular form (question) and a plural form (questions). Tokens #5, #6, and #7 can be any token.

The value attribute in <mi> element uses the lex variable instead of the str variable. It intends to create facet values with no distinction, for example, between “I have a question about Text Mining” and “I have a question about text mining.”

After adding this rule, you must apply them and verify the results again. Rule development is an iterative process. Start with a small set of rules. Then check the results to make sure that they work as expected. Continue to enhance the rules. The rules can be iteratively improved.

Sample rule: Capturing people who are unable to do something

Next you want to understand what customers cannot do. For example, you might look for question-type phrases that are expressed indirectly, such as “I cannot install the application” or “I could not find a catalog.” You first start with the simple rule shown in Example 7-9.

> **Cannot to can not:** If a sentence contains “cannot”, the document processor breaks it down into “can” and “not”.

Example 7-9   Simple rule to capture an expression “someone cannot do something”

```
<mi category="$.voc.question" value="Unable to ${3.lex} ${4.lex}
${5.lex}">
  <w id="1" lex="can"/>
  <w id="2" lex="not"/>
  <w id="3"/>
  <w id="4"/>
  <w id="5"/>
</mi>
```

After applying this rule, you confirm that the example sentences are captured as expected. However, this rule also captures system error messages, such as “Process cannot be terminated in.” For your purposes, you want to distinguish between question-type phrases and error messages. We change the rule as shown in Example 7-10 to accommodate this case.

Example 7-10 Improved rule to capture an expression “someone cannot do something”

```
<mi category="$.voc.question" value="Unable to ${3.lex} ${5.lex}">
  <w id="1" lex="can"/>
  <w id="2" lex="not"/>
  <w id="3" lex="!be"/>
  <w id="4"/>
  <w id="5" pos="noun"/>
</mi>
```

Token #3 uses the lex constraint with the logically not (!) operator. It intends not to match phrases that are in the passive voice. Also we use the part-of-speech constraint in Token #5. The last token has to be a noun. The *pos* attribute stands for part-of-speech. The Pattern Matcher annotator supports 11 parts of speech:

- adjective
- adposition
- adverb
- conjunction
- determiner
- interjection
- noun
- numeral
- pronoun
- residual
- verb

Consider, this sentence, for example:

```
Ah, product #200 is great but really expensive for me!
```

The document processor categorizes tokens into the corresponding part of speech (Table 7-1). You can write a rule that pertains to a particular part of speech.

Table 7-1 Part-of-speech values

Part-of-speech	Token
pronoun	me
verb	is
noun	product
adjective	great expensive
adverb	really
adposition	for
interjection	Ah
conjunction	but
determiner	the
numeral	200
residual	, # !

Token #4 can be anything, such as “a” or “the.” This token is not important in categorizing the common phrase and is, therefore, no longer part of a keyword value template.

Then, you must apply the rule again, and check the results. The rule captures the example sentences and ignores the system error messages as expected. However, the rule might not capture sentences such as “I cannot open files.” If you must capture such sentences as this example, you can add a more generic rule (Example 7-11) in addition to the previous one.

Example 7-11 Additional rule to capture common phrases

```
<mi category="$.voc.question" value="Unable to ${3.lex}">
  <w id="1" lex="can"/>
  <w id="2" lex="not"/>
  <w id="3" lex="!be"/>
</mi>
```

Example 7-10 on page 317 and Example 7-11 can apply to the same sentence.

Sample rule: Capturing negative phrases

To capture the mood of customers, you might want to create rules that are associated with negative phrases. Typical phrases that customers use when they are unhappy or unsatisfied are “... not happy with something” or “... not satisfied with something.” To capture this type of expression, you can use the rules shown in Example 7-12.

Example 7-12 Rules to capture negative phrases

```
<mi category="$.voc.negative-phrase" value="Not ${2.lex} with
${4.lex}">
  <w id="1" lex="/(not$)|(isnt$)|(arent$)/"/>
  <w id="2" lex="/^((happy)|(satisfy)|(satisfactory))$/"/>
  <w id="3" lex="with"/>
  <w id="4" pos="noun"/>
</mi>
<mi category="$.voc.negative-phrase" value="Not ${2.lex} with
${5.lex}">
  <w id="1" lex="/(not$)|(isnt$)|(arent$)/"/>
  <w id="2" lex="/^((happy)|(satisfy)|(satisfactory))$/"/>
  <w id="3" lex="with"/>
  <w id="4" pos="/^((determiner)|(adjective))$/"/>
  <w id="5" pos="noun"/>
</mi>
```

“isnt$” and “arent$” in the token #1 capture “isnt” and “arent” that are denoted literally in a sentence. For example, the rules in Example 7-12 capture a sentence, such as “this is not satisfactory” and “this isn’t satisfactory.” They also capture a sentence such as “this isnt satisfactory.”

The difference between the first rule and second rule is if there is a determiner or an adjective before a noun.

7.4.5 Validation and maintenance

Similar to the Dictionary Lookup annotator, you can check the result of your custom text analysis rules in the Facets view of the text miner application. One alternative to validate the results is to use Real-time Natural Language Processing (NLP) capability, which is useful for evaluating rules interactively. For more information, see 11.3.1, “Real-time NLP” on page 476.

Develop your rules iteratively. Do not try to stretch a rule to make it solve every problem. Complicated rules make maintenance difficult. It is important to keep rules short and simple.

7.5 Preferred practices

Text mining and gaining insight into your content is an iterative process. We developed the following practices based on our field experiences:

- Start your analysis with a small collection.
- Follow the normal procedure several times:
  - Crawl, parse, index, and inspect the content by using the text miner application.
  - Define new dictionaries, and inspect the content by using the text miner application.
  - Define new pattern rules, and inspect the content by using the text miner application.
  - When you discover the need for more sophisticated analysis, use either of the following tools:
    - The IBM Classification Module. See Chapter 9, “Content analysis with IBM Classification Module” on page 357.
    - The IBM LanguageWare Resource Workbench. See 11.2.1, “LanguageWare Resource Workbench” on page 454.

8

Discovering insight with terms of interest and document clustering

Massive amount of textual data usually contains a great wealth of information. A skillful analyst might discover a spectrum of insight by analyzing various facets from this data. However, these invaluable insights might not reveal themselves if you do not select the right facets with their corresponding keywords. IBM Content Analytics provides the term of interest feature that helps you identify these crucial keywords for valuable content analysis.

Another Content Analytics feature, document clustering, helps you detect clusters from data. The resulting clusters can be used to categorize the entire collection, improving your search results or helping you to narrow down the set of documents for detail analysis. Document clustering uses the IBM Classification Module statistical classification algorithm. This chapter describes the usage.

This chapter includes the following sections:

- The power of dictionary-driven analytics
- Terms of interest
- Document clustering

8.1 The power of dictionary-driven analytics

To acquire valuable insights from your data, it is important to identify the appropriate keywords and analyze their deviations and changes. Aimless analysis of facets with a broad range of values, such as all the nouns in your data collection, is not a good approach because the amount of irrelevant terms (or noise) often masks the most important insights. Therefore, it is essential to select appropriate keywords for each specific purpose of analysis. It is also necessary to register them into a dictionary and pattern matching rules that are associated with their proper facets.

To help users identify these important keywords, Content Analytics Version 2.2 provides automatic identification of candidate terms of interest. This feature is based on Data Oriented Lexical Creation Engine (DOLCE) technology that automatically derives candidate terms of interest from the textual content of your collection. For more details about this function, see 8.2, “Terms of interest” on page 326.

8.1.1 Multiple viewpoints for analyzing the same data

As explained in 3.2.1, “Setting the objectives of the analysis” on page 53, an important step for taking advantage of Content Analytics is to set appropriate objectives for analysis. That is, match what you want to do with the data and what the data allows you to do. In our experience, even when the data does not provide the answers you need, it often discovers valuable insights that are unexpected. A dictionary plays a key role in customizing Content Analytics for each of your analysis objectives.

Analysis of complaint data for problem identification

Consider an example of analyzing complaint data about cars that consists of problem reports from various drivers. In fact, such data is maintained and made public in many countries including the US, Japan, China, and France. This type of data typically consists of a textual description of each problem. It is accompanied by structured information, such as the report date, model of car, name of automotive company, and the area where the driver lives.

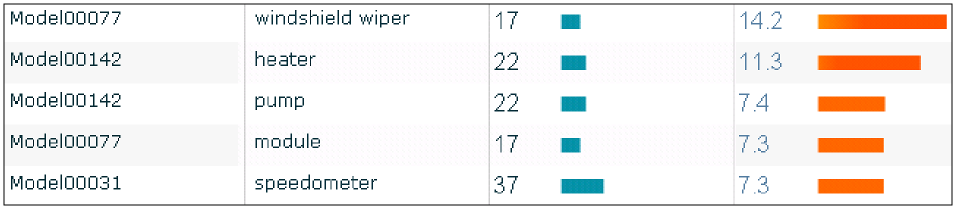
Many of the customer contact records, containing voice-of-customer (VOC) data, share some essence with this complaint data about cars in terms of the textual description associated with various structured information.

A major use of complaint data for an automotive company is to identify critical defects that need fixing hopefully in their early stages. For this analysis, it is important to analyze the type of problems that occurred with which components.

Therefore, for this analysis, the following facets might be appropriate:

- The Problem facet, which consists of keywords that describe the nature of the problem such as “leak,” “crack,” “fire,” and “blow”
- The Component facet, which consists of keywords that describe the car components that are involved such as the “brake,” “engine,” “transmission,” and “steering wheel”

By applying correlation analysis based on these two facets using the Facet Pairs view, you can identify notable defects for a specific car model compared to other models. For example, in Figure 8-1, which shows the correlation between car models and components, Model00077 has a strong correlation with the windshield wiper.



Model00077	windshield wiper	17	14.2
Model00142	heater	22	11.3
Model00142	pump	22	7.4
Model00077	module	17	7.3
Model00031	speedometer	37	7.3

Figure 8-1 Facet Pairs view for correlation analysis between car models and components

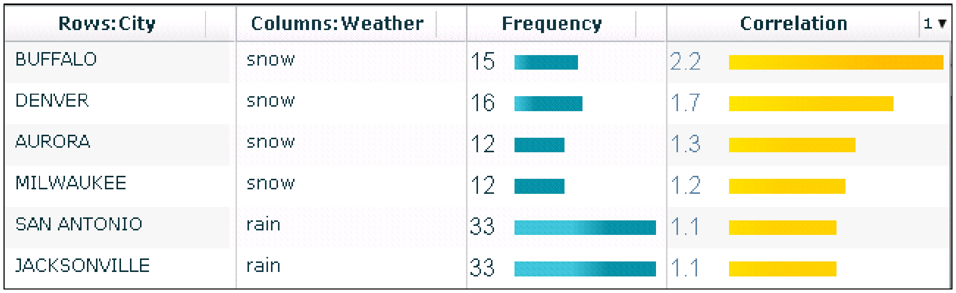
The correlation index indicates that this model tends to have a problem with the wiper approximately 14 times higher than other models. By focusing on the 17 reports for Model00077 that describe a problem with the windshield wiper, you can check the Problem facet view to analyze the kind of problem that is typically reported on a wiper of this model. In this case, the majority of the 17 reports claims a similar phenomenon. Their windshield wipers did not always work when the switch was turned on, and they did not always stop when the switch was turned off.

Analysis for additional insights from various viewpoints

The use of complaint data about cars is not limited to the analysis of car problems. In the complaint data, each problem is usually described in context, such as who is involved and in which circumstances.

Weather

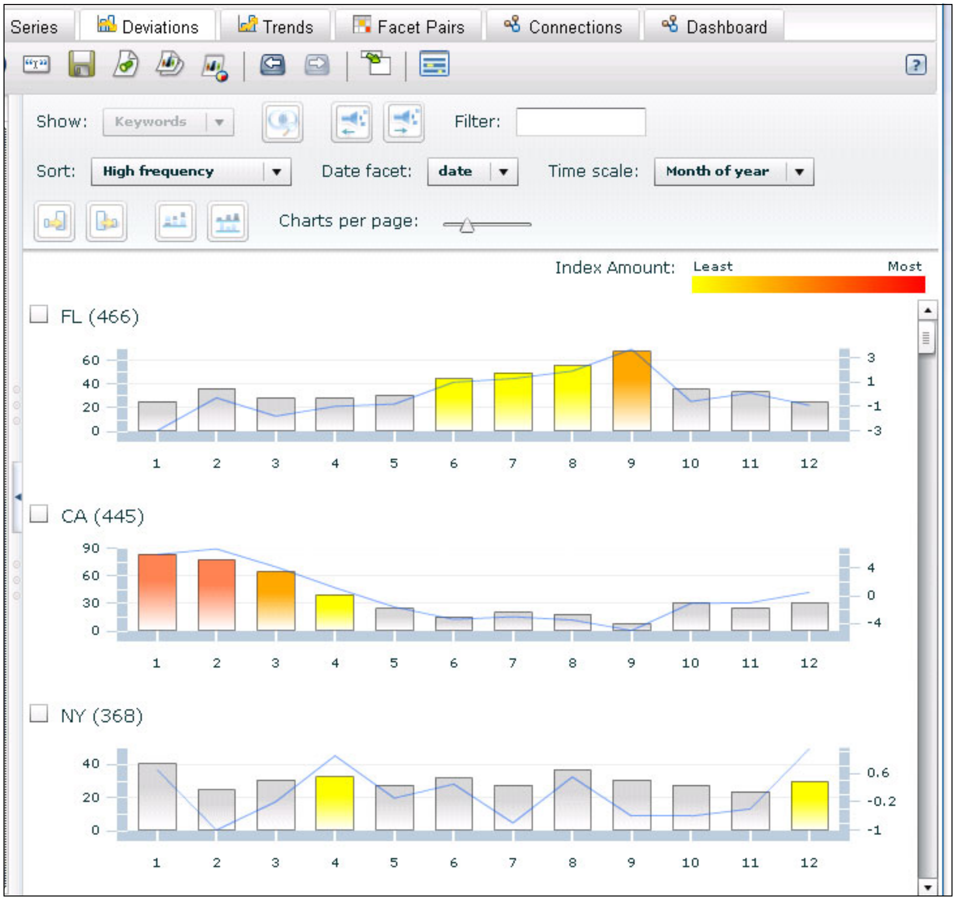
Consider a case where most of the data contains information about the weather conditions. By defining a Weather facet that consists of keywords such as “rain” and “snow,” you can analyze which cities are strongly associated with what type of weather. The example shown in Figure 8-2 on page 324 indicates that Buffalo has a stronger association with snow compared to other cities.



Rows:City	Columns:Weather	Frequency	Correlation
BUFFALO	snow	15	2.2
DENVER	snow	16	1.7
AURORA	snow	12	1.3
MILWAUKEE	snow	12	1.2
SAN ANTONIO	rain	33	1.1
JACKSONVILLE	rain	33	1.1

Figure 8-2 Facet Pair view for analyzing the correlation between city and weather

By analyzing the distribution of states with the month of the year in the Deviation view after focusing on the data with rain, you can see the rainy season for each state as revealed in Figure 8-3.



Show: Keywords

Filter:

Sort: High frequency

Date facet: date

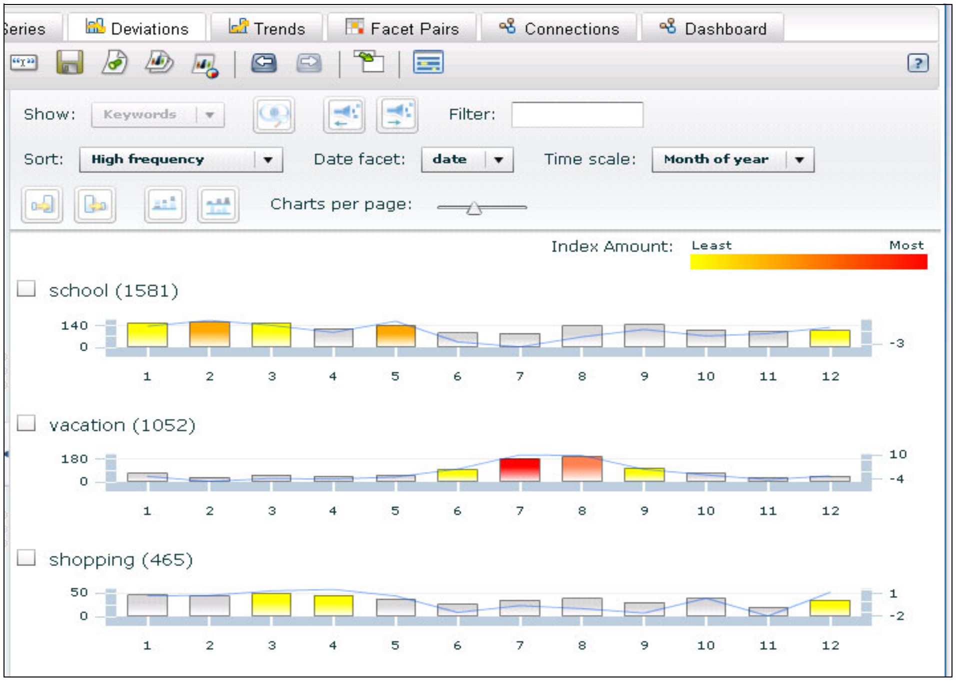
Time scale: Month of year

Charts per page:

Figure 8-3 Deviations view for analyzing rain data by state

Activities and families

Some of the data also contains information related to driver activities such as shopping, vacation, and school. Figure 8-4 indicates the high season within a year for each activity.



Show: Keywords

Filter:

Sort: High frequency

Date facet: date

Time scale: Month of year

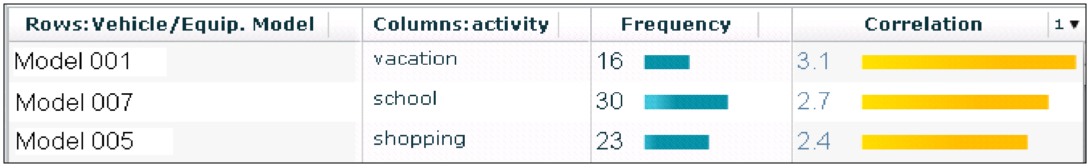
□ school (1581)

□ vacation (1052)

□ shopping (465)

Figure 8-4 Deviations view for analyzing monthly activities based on user-defined keywords

Figure 8-5 indicates which models are highly associated with each activity.



Rows:Vehicle/Equip. Model	Columns:activity	Frequency	Correlation 1 ▾
Model 001	vacation	16	3.1
Model 007	school	30	2.7
Model 005	shopping	23	2.4

Figure 8-5 Facet Pair view for analyzing the correlation between model and activity

Likewise, much of the data also contains information about families such as wife, kids, and parents. With this information, you can also analyze which cars are correlated to which family type and which activity.

Extending use cases

By expanding the viewpoints with new facets, you can extend use cases. In the complaint data about cars, typical usage of car models might be identified, and such information might be valuable for product developments and target marketing. Analysis of environments might help manufacturers find good places and seasons for field-test environments.

Moreover, information from complaint data about cars can benefit car manufacturers, their dealers, and the following groups:

- Drivers to identify potential problems of their own cars
- Used-car dealers to estimate car conditions by model and manufacturing year
- Insurance companies to estimate the risk of each car

8.2 Terms of interest

As described in 8.1, "The power of dictionary-driven analytics" on page 322, a large amount of textual data can provide various insights. However, most novice users tend to analyze data without a customized dictionary. Novice users also tend to perform aimless analysis of their data by using a broad range of expressions such as all the nouns that might not lead to any valuable results.

To help users gain valuable insight, IBM Content Analytics Version 2.2 provides automatic identification of candidate terms of interest. This feature is based on DOLCE technology that automatically derives candidate terms of interest from the textual content of your collection.

To enable terms of interest when creating your collection, follow these steps:

1. Click the **Create Collection** button (Figure 8-6).

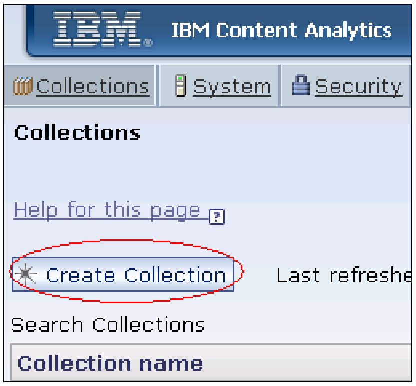
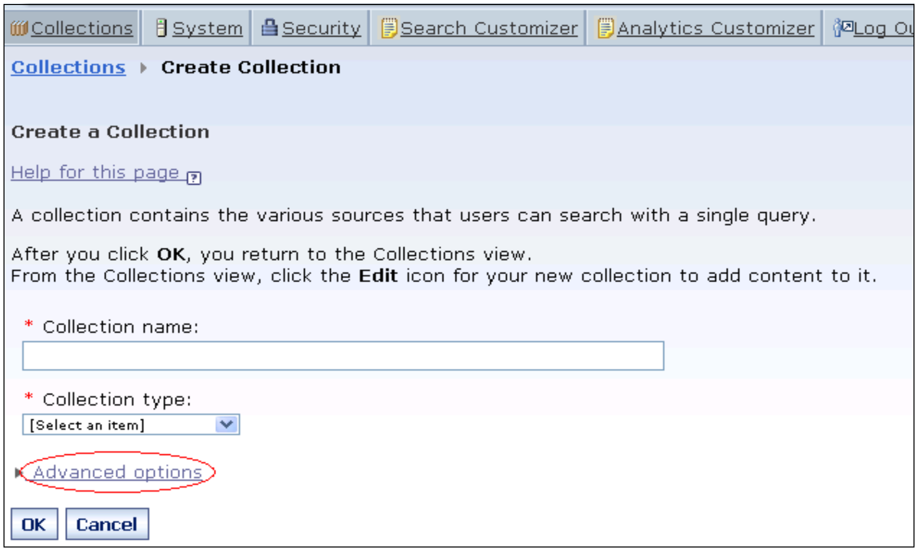


Figure 8-6 Create Collection button on the System Administration Console

2. In the Create a Collection panel (Figure 8-7), click **Advanced options**.



Collections ▸ Create Collection

Create a Collection

Help for this page

A collection contains the various sources that users can search with a single query.

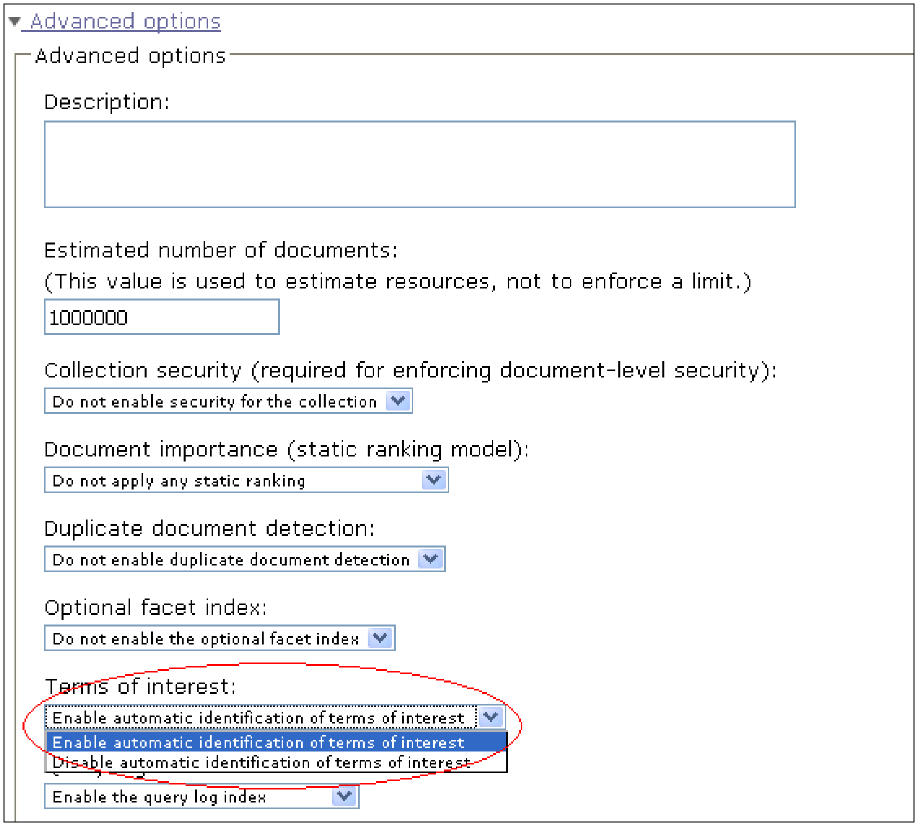
After you click **OK**, you return to the Collections view.
From the Collections view, click the **Edit** icon for your new collection to add content to it.

* Collection name:

* Collection type:
[Select an item]

Figure 8-7 Selection of Advanced options

3. Under Advanced options (Figure 8-8), for Terms of interest, select **Enable automatic identification of terms of interest** to activate the terms of interest feature.



▼ Advanced options

Advanced options

Description:

Estimated number of documents:
(This value is used to estimate resources, not to enforce a limit.)
1000000

Collection security (required for enforcing document-level security):
Do not enable security for the collection

Document importance (static ranking model):
Do not apply any static ranking

Duplicate document detection:
Do not enable duplicate document detection

Optional facet index:
Do not enable the optional facet index

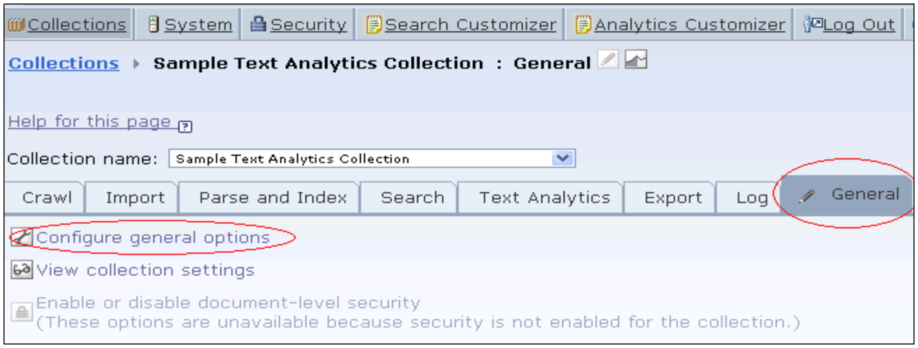
Terms of interest:
Enable automatic identification of terms of interest
- Enable automatic identification of terms of interest
- Disable automatic identification of terms of interest

Enable the query log index

Figure 8-8 Selecting 'Enable automatic identification of terms of interest'

For collections that are already created, you can enable this option by following these steps:

1. On the **General** tab, click **Configure general options** (Figure 8-9).

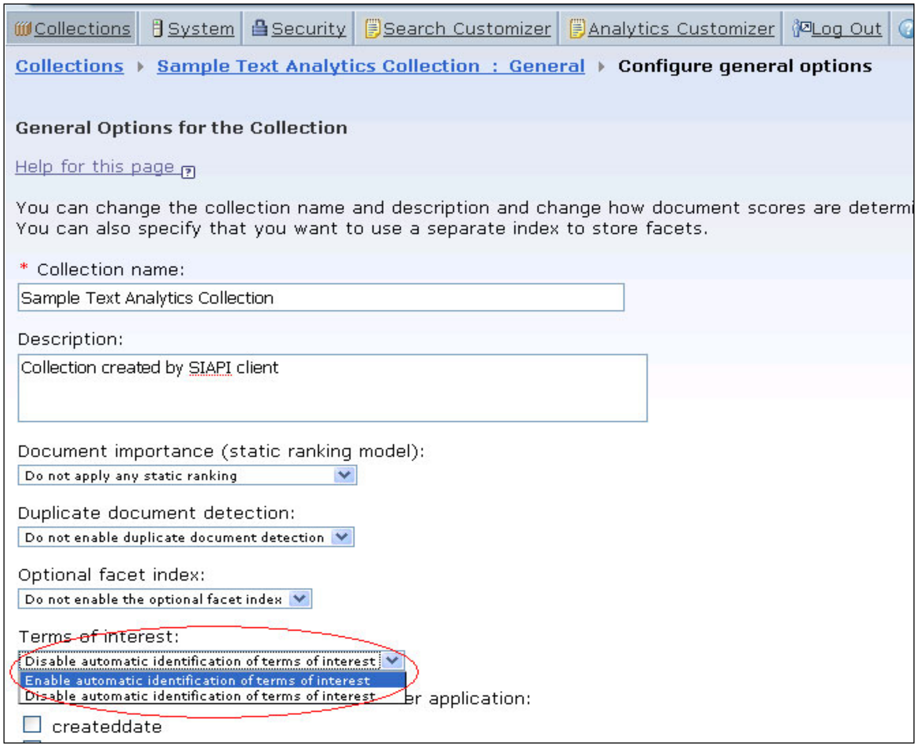


Collection name: Sample Text Analytics Collection

Enable or disable document-level security
(These options are unavailable because security is not enabled for the collection.)

Figure 8-9 Clicking 'Configure general options' on the General tab

2. In the General Options for the Collection window (Figure 8-10), for Terms of interest, select **Enable automatic identification of terms of interest**.



General Options for the Collection

You can change the collection name and description and change how document scores are determi
You can also specify that you want to use a separate index to store facets.

* Collection name:
Sample Text Analytics Collection

Description:
Collection created by SIAPI client

Document importance (static ranking model):
Do not apply any static ranking

Duplicate document detection:
Do not enable duplicate document detection

Optional facet index:
Do not enable the optional facet index

Terms of interest:
Disable automatic identification of terms of interest
Enable automatic identification of terms of interest
Disable automatic identification of terms of interest

er application:

□ createddate

Figure 8-10 Selecting 'Enable automatic identification of terms of interest'

After enabling automatic identification of terms of interest, perform the steps in “Resource deployment” on page 135 and “Rebuilding the full index” on page 137.

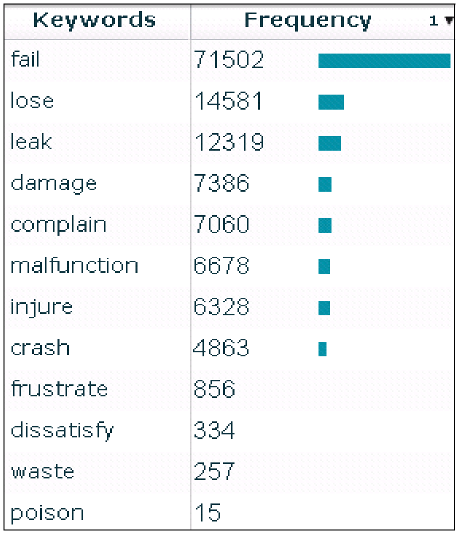
8.2.1 Basic algorithm for identifying terms of interest

Problem identification has been one of the most promising use-case scenarios for Content Analytics because it usually leads to action for reducing and preventing a problem.

Predicates in Terms of Interest

In the complaint data about cars, problems are often described with predicates (verbs), such as “fail,” “stop,” and “leak.” To recognize complaints involving such predicates, you might prepare dictionary and pattern matching rules to be used by Content Analytics to identify these unfavorable situations such as “fail,” “damage,” and “complain.”

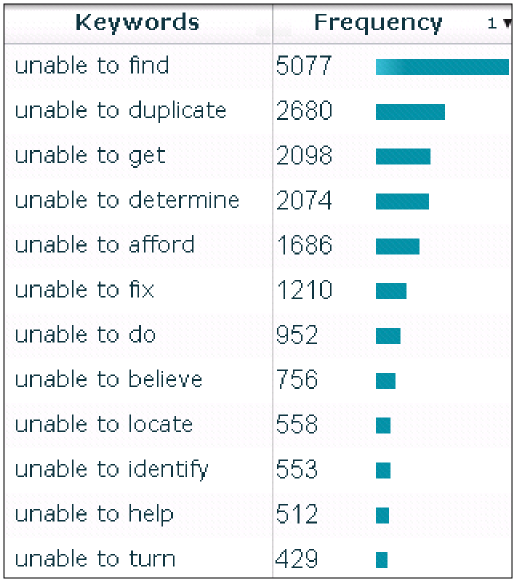
Figure 8-11 shows the distribution of general problem-type verbs in the nearly 620,000 car complaint records. However, many of these verbs do not indicate specific problems but rather indicate the existence of a potential problem. Moreover, specific problems usually depend on the domain context of the data. This domain dependence makes it impractical to predefine such problem expressions for diverse domains.



Keywords	Frequency 1▼
fail	71502
lose	14581
leak	12319
damage	7386
complain	7060
malfunction	6678
injure	6328
crash	4863
frustrate	856
dissatisfy	334
waste	257
poison	15

Figure 8-11 Distribution of predefined general problem verbs within car complaints

Another approach for capturing problem expressions is to use certain patterns such as verbs followed by “cannot,” “failed to,” and “not able to” as shown in Figure 8-12. Although this approach might capture a wider variety of expressions, the frequency of many of such expressions is not very high, and the uses of low frequency expressions are not suitable for analysis of deviations and changes of trends.



Keywords	Frequency 1▼
unable to find	5077
unable to duplicate	2680
unable to get	2098
unable to determine	2074
unable to afford	1686
unable to fix	1210
unable to do	952
unable to believe	756
unable to locate	558
unable to identify	553
unable to help	512
unable to turn	429

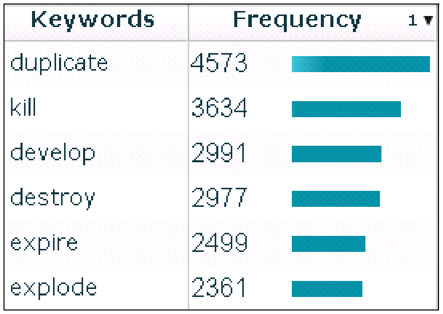
Figure 8-12 Distribution of expressions with predefined patterns within car complaints

A new function of Content Analytics Version 2.2 is the automatic identification of terms of interest. Terms of interest are determined by using the nature of problem expressions that tend to co-occur with specific adverbial expressions such as “suddenly” and “often.” However, they do not co-occur with specific adverbial expressions such as “correctly” and “normally.”

For example, problems are typically reported in textual data with expressions such as “my machine suddenly freezes,” indicating that “freeze” is the problem. Yet, non-problem expressions can be also modified by “suddenly” such as in “it suddenly worked,” indicating a recovery from some previous problem. On the contrary, problem expressions are seldom modified by “correctly” and “normally.” For example, “It freezes correctly” sounds odd if “freeze” indicates a problem.

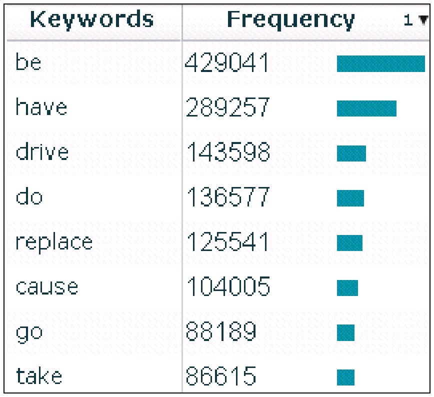
Thus, by calculating the co-occurrence ratio of verbs with such adverbial expressions within an entire collection of data, Content Analytics can identify candidate problem expressions automatically. It can also list them in the Predicate subfacet under the Terms of Interest facet. Because such adverbial expressions are limited and their concepts are language independent, Content Analytics provides this function for all supported languages.

Predicate in Terms of Interest usually contains keywords that indicate potential problems (Figure 8-13) compared to Verbs in Part of Speech that lists all verbs without focusing on problem areas (Figure 8-14).



Keywords	Frequency 1 ▼
duplicate	4573
kill	3634
develop	2991
destroy	2977
expire	2499
explode	2361

Figure 8-13 Predicate in Terms of Interest within a collection of car complaints



Keywords	Frequency 1 ▼
be	429041
have	289257
drive	143598
do	136577
replace	125541
cause	104005
go	88189
take	86615

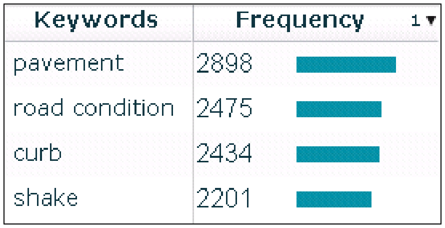
Figure 8-14 Verbs in Part of Speech within a collection of car complaints

In general, this algorithm works well for a collection that contains a large number of records with problem descriptions. For a collection with a relatively small number of records or with textual data that does not contain many problem descriptions, this function might not identify any terms of interest. Also the facet might be sparsely populated. In particular, if none of the records contain “suddenly,” “often,” “sometimes,” or “frequently” in the case of English texts, this function will not identify any terms of interest, and the Predicate subfacet under Terms of Interest facet will be empty.

Entities in Terms of Interest

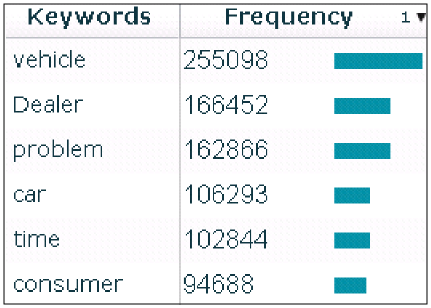
When Predicates in Terms of Interest are determined as candidate expressions, Content Analytics looks for nouns that tend to be associated with these predicates. These nouns are typically their subjects and direct objects and can be classified as candidates of entities that relate to the problem. As a result, identified nouns are listed in the Entity subfacet under the Terms of Interest facet.

Entity in Terms of Interest usually contains keywords that relate to a set of problems (Figure 8-15) as compared to the Nouns in Part of Speech that lists all nouns without bringing attention to specific problem areas (Figure 8-16).



Keywords	Frequency 1 ▼
pavement	2898
road condition	2475
curb	2434
shake	2201

Figure 8-15 Entity in Terms of Interest within a collection of car complaints



Keywords	Frequency 1 ▼
vehicle	255098
Dealer	166452
problem	162866
car	106293
time	102844
consumer	94688

Figure 8-16 Nouns in Part of Speech within a collection of car complaints

Keywords of “predicate” in “terms of interest” represent candidate problems. However, keywords in “entity” in “terms of interest” represent potential entities that relate to problem such as the cause of the problem and targets of the problem.

Similar to the case of Predicate, this algorithm generally works well for a collection that contains a large number of records with problem descriptions. For a collection with a relatively small number of records or with textual data that does not specify any problem descriptions, this function might not identify any terms of interest.

In particular, because keywords in the Entity facet are derived from keywords in the Predicate facet, the values in the Entity facet are empty if the values of Predicate facet are empty. If none of the records contain “suddenly,” “often,” “sometimes,” or “frequently” in English texts, this function does not identify any terms of interest, and both the facets will be empty.

8.2.2 Limitations in using automatic identification of terms of interest

The function for identifying terms of interest can help users to identify valuable insights through Content Analytics. The algorithm used to identify terms of interest is based on the statistical distribution of keywords associated with a set of specific adverbs. Therefore, the quality of the result varies significantly depending on the textual data within each collection. In particular, it requires a reasonable amount of predicate descriptions associated with “suddenly,” “often,” “sometimes,” “frequently,” “correctly,” “firmly,” “normally,” “properly,” and “securely,” and corresponding expressions for other languages.

In addition, because the algorithm is designed to be robust and language independent, the presence of noise (non-problem terms) is unavoidable. Therefore, treat terms of interest as *candidate* terms of interest.

Based on our experiments, this algorithm works well for a collection that consists of a million records or more and that contains problem descriptions within a small domain where the use of each term is relatively consistent. However, it is important to understand the limitation of the terms of interest function. Terms of interest might not work well for the following types of collection:

- A collection with a small number of records
- A collection without problem descriptions
- A collection with textual content in a diverse domain where many of the keywords are polysemous with multiple meanings

  For example, both use of “freeze” for “refrigerate” and “halt” are observed in textual data within the same collection.

8.2.3 Preferred use of terms of interest identified automatically

To help users to gain valuable insight through Content Analytics, the keywords listed in the Terms of Interest facet might be used directly for problem detection and for correlation analysis to identify specific problems. In addition, the list of keywords provides good candidates to include in a dictionary.

Terms of interest for problem detection

Not all terms in Predicate of Terms of Interest are relevant to a specific problem. However, going through the list of keywords in Predicate through Facet view can lead to actionable insights.

For example, through the analysis of a collection of customer satisfaction surveys, the keyword "change" was listed in the Facet view of Predicate under Terms of Interest. By focusing on the textual data containing "change," we saw that a sudden change and frequent changes of their customer representative had a negative impact on customer satisfaction. This kind of insight leads to action, resulting in a modification to the customer relationship management strategy.

Therefore, looking at each keyword in Predicate of Terms of Interest can lead to the identification of noteworthy problems.

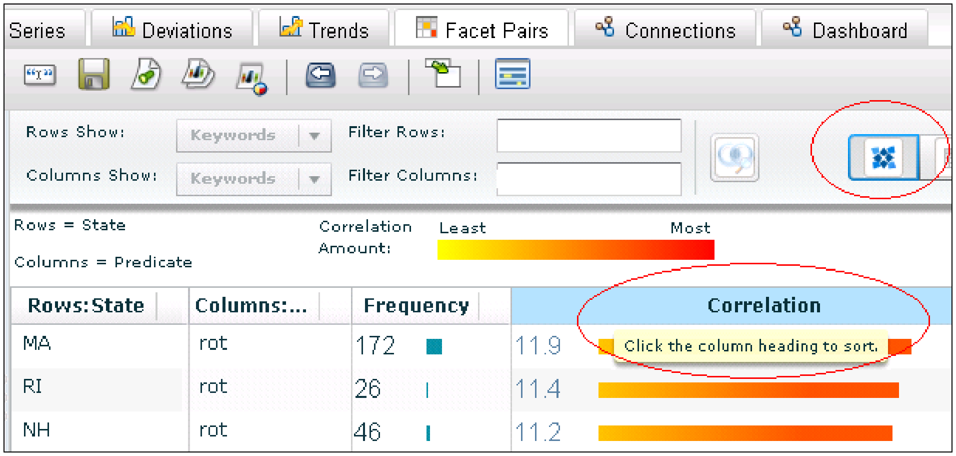
Terms of interest for correlation analysis

The skill for acquiring valuable and actionable insights by using Content Analytics might require practice. It requires good sense, imagination, deep domain knowledge, and patience. However, you can improve this skill based on your experience. By using terms of interest, you can acquire valuable and actionable insights relatively easily.

When analyzing your collection, follow these steps:

1. Open the Facet view and select **Predicate under Terms of Interest**.
2. If you see a reasonable number of keywords listed and some of them seem to indicate a potential problem, open the Facet Pairs view by selecting **Predicate** under Terms of Interest in the Column.

3. Select other facets as the Rows by sorting the Table view with Correlation (Figure 8-17).



Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns:

Rows = State

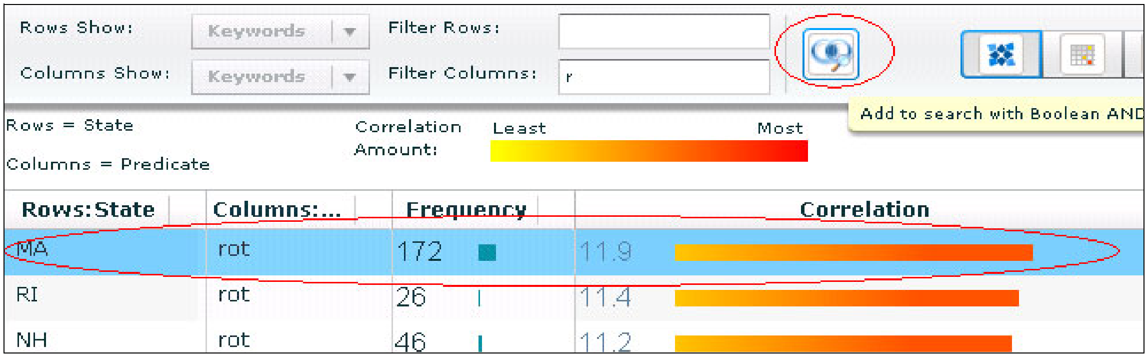
Columns = Predicate

Correlation Amount:

Rows:State	Columns:...	Frequency	Correlation
MA	rot	172	11.9
RI	rot	26	11.4
NH	rot	46	11.2

Figure 8-17 Facet Pairs view sorted by correlation

4. If you find any noticeable correlation pairs, select the pair (Figure 8-18).



Rows Show: Keywords

Columns Show: Keywords

Filter Rows:

Filter Columns: r

Rows = State

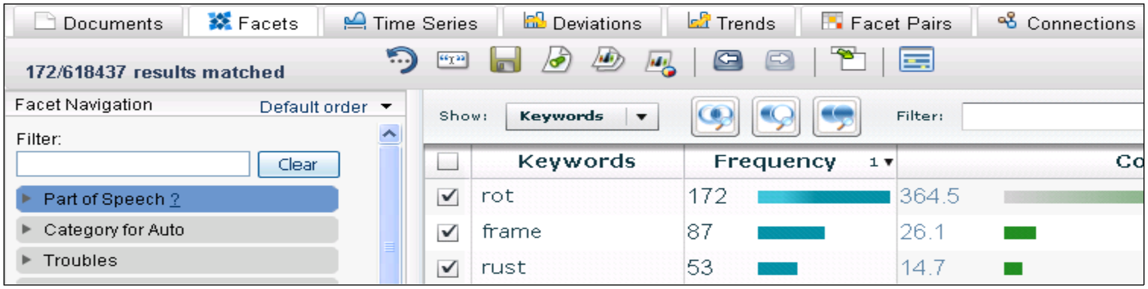
Columns = Predicate

Correlation Amount:

Rows:State	Columns:...	Frequency	Correlation
MA	rot	172	11.9
RI	rot	26	11.4
NH	rot	46	11.2

Figure 8-18 Selecting a specific pair to focus on the data set associated with both keywords

5. Check the details of the data with the Facet view by selecting **Parts of Speech** and sorting the list by Correlation (Figure 8-19).

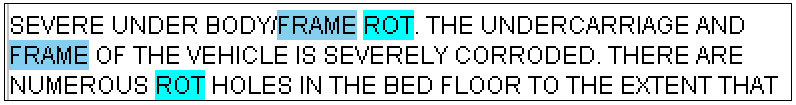


172/618437 results matched

□	Keywords	Frequency 1 ▾	Co
☑	rot	172	364.5
☑	frame	87	26.1
☑	rust	53	14.7

Figure 8-19 Analysis of features of the data focused at the Facet Pairs view

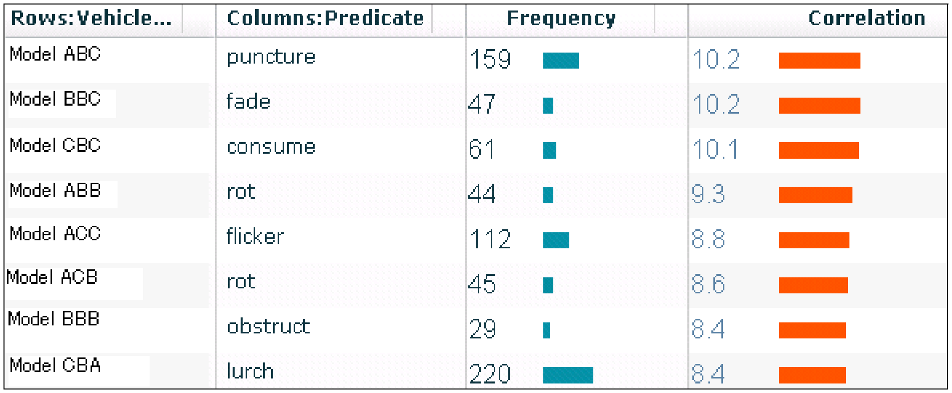
In this example, the data indicates “rot” problems (with “frame” and “rust”) are typical in the northern eastern states of the US (New Hampshire, Road Island, and Massachusetts). You can verify this assumption by selecting the keywords “rot,” “frame,” and “rust,” and checking the Document view as shown in Figure 8-20.



SEVERE UNDER BODY/FRAME ROT. THE UNDERCARRIAGE AND FRAME OF THE VEHICLE IS SEVERELY CORRODED. THERE ARE NUMEROUS ROT HOLES IN THE BED FLOOR TO THE EXTENT THAT

Figure 8-20 Keywords identified from the Facet Pairs view and Facet view

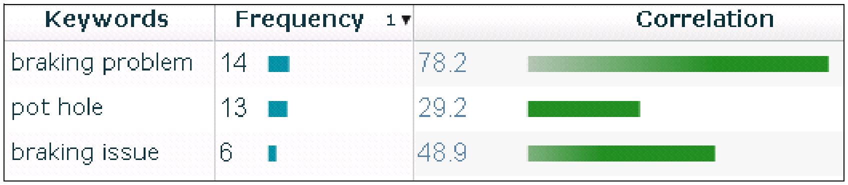
In general, selection of product names often leads to valuable and actionable insights because it indicates problems that are specific to some products. For example, in the analysis of the complaint data about cars, you can find product-specific problems similar to those problems shown in Figure 8-21. Figure 8-21 shows a Facet Pairs view with facets from Model and Predicate in Terms of Interest.



Rows:Vehicle...	Columns:Predicate	Frequency	Correlation
Model ABC	puncture	159	10.2
Model BBC	fade	47	10.2
Model CBC	consume	61	10.1
Model ABB	rot	44	9.3
Model ACC	flicker	112	8.8
Model ACB	rot	45	8.6
Model BBB	obstruct	29	8.4
Model CBA	lurch	220	8.4

Figure 8-21 Car models and their predicates for potential car complaint problems

As a result, you might find that many drivers of Model CBA reported that the car lurches when braking around pot holes as illustrated in Figure 8-22. The keywords shown in Figure 8-22 are shown in the Facet Pairs view through the Facet view for Phrase Constituent → Noun Phrase → Noun Sequence.



Keywords	Frequency 1▼	Correlation
braking problem	14	78.2
pot hole	13	29.2
braking issue	6	48.9

Figure 8-22 Keywords as potential problems associated with Model CBA and lurch Predicate

It is important to try the various facets in step 3 on page 336 with patience and imagination.

Terms of interest as candidates for dictionary creation

As explained in 8.1, “The power of dictionary-driven analytics” on page 322, a customized dictionary and pattern matching rules enrich your analysis and often lead you to better insights. It is always easier and better to make a dictionary and pattern matching rules from a meaningful list of candidate keywords instead of working from scratch.

The list of keywords in the Predicate of Terms of Interest facet is a good resource for creating pattern matching rules for a facet that might represent potential problems. To deal with a conjugated form, register predicates as pattern matching rules rather than registering them in a dictionary where terms are treated as nouns.

Compared to all the verbs in a collection, Predicates from Terms of Interest help narrow down the keywords that indicate potential problems. As an example, you can compare the keywords for all verbs in a collection in Figure 8-14 on page 332 with the keywords that indicate some problems in Figure 8-13 on page 332. Therefore, the use of keywords in Predicate of Terms of Interest makes the creation of pattern matching rules (as shown in 7.2.2, “Scenario 2: Using custom text analysis rules to discover trouble-related calls” on page 291) productive and effective.

It is important to understand that the Predicate of Terms of Interest might not contain all of the keywords that indicate potential problems. That is, pattern matching rules based on the keywords in just the Predicate facet might not be a complete set of pattern matching rules. However, creation of a dictionary and pattern matching rules is an iterative procedure. You must create a dictionary and pattern matching rules for a meaningful facet quickly. You must also apply the dictionary and pattern matching rules to your collection, and test how they lead to valuable insights.

After you understand the value of the dictionary and pattern matching rules for analysis, go through a longer list of keywords in a facet under Part of Speech, such as Noun and Verb, to enrich the dictionary and pattern matching rules.

Like the list of keywords in Predicate, the list of keywords in Entity of Terms of Interest is a good resource for creating dictionaries. In the case of keywords for the Entity facet, you might want to define multiple facets, such as “parts” or “components” in the car complaint scenario that are damaged and “causes” that create the problem.

Again, do not treat the list of keywords in the “entity” facet as a complete list. It might contain noise that is not relevant to any of the facets that defined. Also, it might not contain all the keywords to be registered under each facet. Therefore, it is important to create a tentative dictionary quickly with minimum effort and apply it to your collection to check the feasibility to acquire such valuable insights.

8.2.4 Efficient and effective creation of dictionary

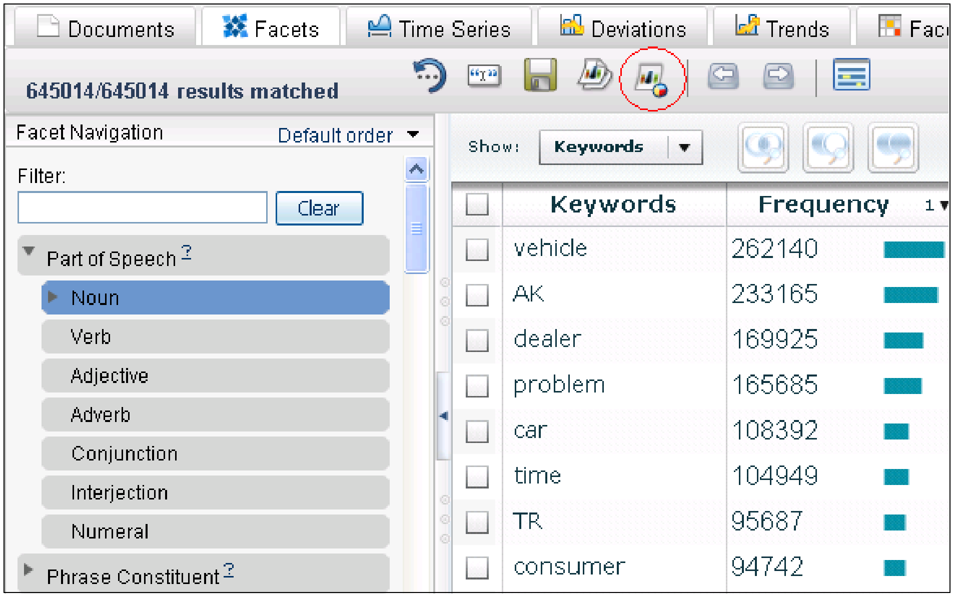
To efficiently and effectively create a dictionary, you start with a list of keywords. You obtain a better list by reviewing terms of interest, using rules and correlation values.

Using a list of keywords for dictionary creation

To create a Content Analytics dictionary, the best resource of terms to be registered in the dictionary and pattern matching rules is the collection that you are working with.

To identify deviations and changes in keyword distribution that lead to valuable insights, the keywords in the dictionary must appear frequently in the textual data of the collection. If none of the terms in the dictionary appear in your data, the dictionary is useless. Therefore, instead of building a dictionary from scratch, always start with a list of keywords that are extracted from Content Analytics.

For this purpose, a list of nouns in the Facet view is a typical candidate list of keywords to be in many facets and in the Noun Sequence facet under the Noun Phase of Phrase Constituent. This list can be easily exported to your machine in the comma-separated value (CSV) format (Figure 8-23 on page 340) so that you can edit it by using a conventional spreadsheet application.



645014/645014 results matched

□	Keywords	Frequency
□	vehicle	262140
□	AK	233165
□	dealer	169925
□	problem	165685
□	car	108392
□	time	104949
□	TR	95687
□	consumer	94742

Figure 8-23 CSV output for nouns listed in the Facets view

However, before downloading the list, change the preferences (by clicking the **Preferences** link as shown Figure 8-24) so that you have a bigger list of keywords to work with.

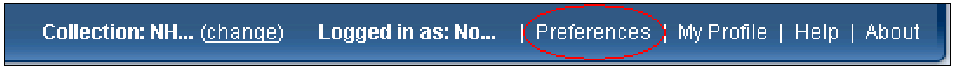
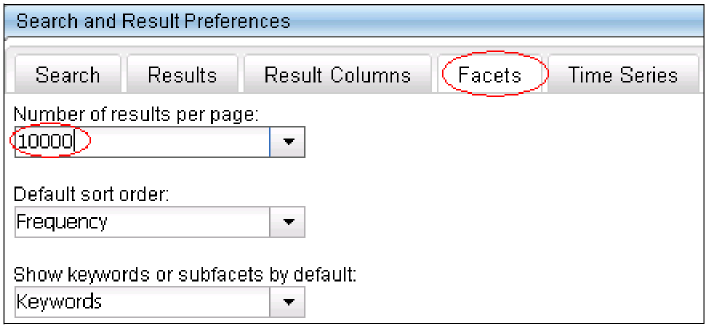


Figure 8-24 Clicking the Preferences link

Then in the Search and Result Preferences window (Figure 8-25), you can type any number instead of selecting a number from the list. The larger number you set for this preference, the longer it takes to see the list in the Facet view. After you download the list, reset the preference to recover a good response time.



Number of results per page:
10000

Default sort order:
Frequency

Show keywords or subfacets by default:
Keywords

Figure 8-25 Setting Preference for Facets view

With this list, make a list of keywords for each facet by editing the file, and configure your dictionary as shown in 7.3.2, “Configuring custom user dictionaries” on page 300. During this procedure, go through the list from higher frequently occurring keywords to lower frequently occurring keywords. Do not resister keywords that appear only once in the entire data. In addition, do not sort the keywords alphabetically unless you are looking for synonyms.

Making a better list of keywords for dictionary creation

A keyword list for facets under Part of Speech and Phrase Constituent is often too general. Therefore, you might have to go through hundreds or thousands of keywords to select tens of keywords to be registered. You might want to make a better list that contains more keywords to be registered within a smaller number of candidates.

Using Terms of Interest

The list of values under Predicate under Terms of Interest is a list for registering pattern matching rules to a facet that indicates potential problems.

The list of Entity under Terms of Interest might be helpful for deciding the types of facets to define. It tends to contain many entities that are involved in problem identification and to contain many keywords to be registered to dictionaries.

Using rules

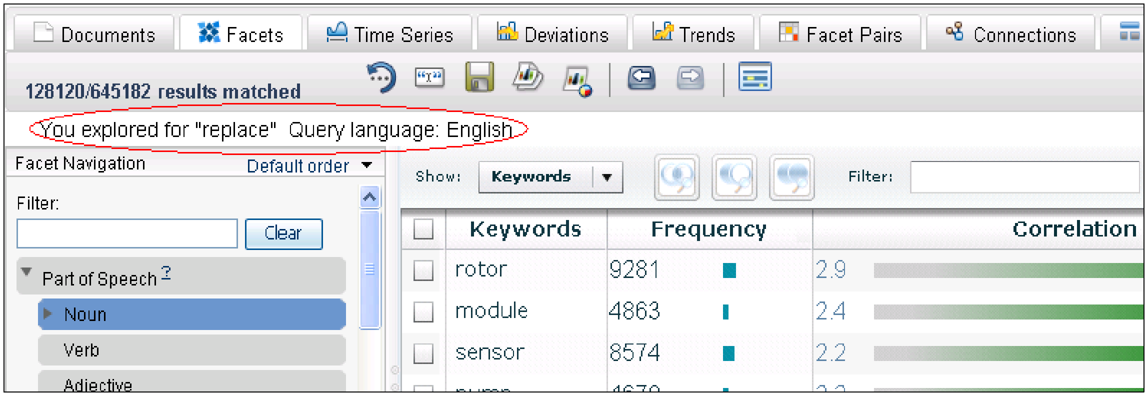
If you define a facet whose keywords might typically appear in a certain context, you can extract keywords by using rules for the Pattern Matcher annotator as explained in 7.4, “Configuring the Pattern Matcher annotator” on page 309.

For example, if you are trying to define a Food facet, the configuration of a rule that extracts nouns that appear right after the verb “eat” might likely capture something that is eatable. In the same manner, the configuration of a rule that extracts nouns that appear right after the verb “replace” might likely capture something that is replaceable.

Using correlations

Instead of using the Pattern Matcher annotator, you can use a combination of searching with relevant expressions and sorting with correlation in the Facets view. With this approach, you can quickly acquire a fairly good list of keywords for dictionary creation.

For example, by making a search with “replace” and list Nouns or Noun Sequences in Facet view by sorting with correlation, you can see many replaceable words as in the lists shown in Figure 8-26 and Figure 8-27 on page 342.

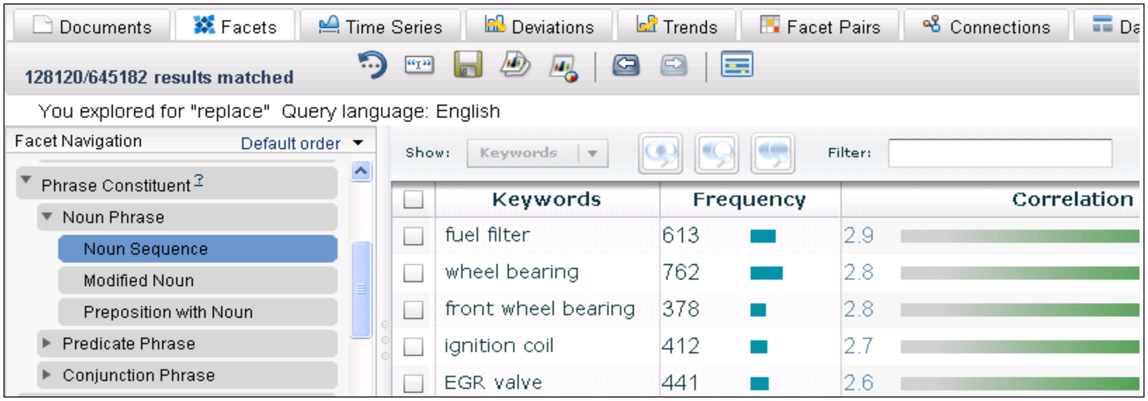


128120/645182 results matched

You explored for "replace" Query language: English

Keywords	Frequency	Correlation
rotor	9281	2.9
module	4863	2.4
sensor	8574	2.2

Figure 8-26 Facets view for the ‘replaceable’ noun



128120/645182 results matched

You explored for "replace" Query language: English

Keywords	Frequency	Correlation
fuel filter	613	2.9
wheel bearing	762	2.8
front wheel bearing	378	2.8
ignition coil	412	2.7
EGR valve	441	2.6

Figure 8-27 Facets view for ‘replaceable’ noun sequence

The result might not be as good as the list generated by using rules. However, it is much easier to make this list, and this method is often worth trying.

8.3 Document clustering

The document clustering functionality in Content Analytics is empowered by the Classification Module clustering algorithms. With document clustering, you can obtain insight quickly into a large data set of unstructured data without setting up dictionaries or defining rules. It provides a potential categorization of your documents. You can use automatic cluster-based categorization to navigate through your content to gain insight into your content.

The categorization is based on a sample content set of 1,000 to 10,000 documents. Document clustering does not require that you install and configure Classification Module. However, the cluster categorization results for document clustering might be less informative or detailed.

To run document clustering, you need to use a large subset of data for Content Analytics to determine meaningful categories. After Content Analytics determines the clusters, you can review, rename, remove, or add clusters. When you are satisfied with the defined clusters, you can apply them to the entire collection. Then, a facet that represents the document cluster results is generated.

The document clustering workflow consists of the following tasks, which are explained in the sections that follow:

1. Setting up document cluster
2. Creating a cluster proposal
3. Refining the cluster results
4. Deploying clusters to a category

8.3.1 Setting up document cluster

To use the document clustering functionality, you must specify that the collection supports document clusters for the categorization type when you create the collection. There are four categorization type options:

- None
- Rule based
- Document clusters
- Rule based and Document clusters

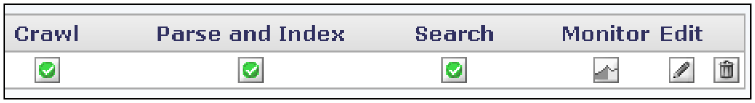
The Document clusters option enables document clustering for the collection. With the Rule-based and Document clusters, you can define rules and perform document clustering for the collection.

After the collection is created, and documents are crawled and indexed, you can perform document clustering. To configure document clustering, follow these steps:

1. From the administration console, click the **Collections** tab.
2. Click **Create Collection**.
3. In the Create a Collection pane, complete the following fields.
   a. For the Collection name field, type `LargeSampleCollection`.
   b. For the Collection type field, select **Text analytics collection**.
   c. For the Categorization type field, select either *Document clusters* or *Rule-based and document clusters* to enable document clustering. For this scenario, select **Document clusters**.

   > **Text analytics collection:** Document clustering can only be enabled for a text analytics collection because it is not available for search collections.

   d. Click **OK**.
4. In the Collections view, click the **Edit** icon (Figure 8-28) for the collection that you want to edit. We click the **Edit** icon that corresponds to the LargeSampleCollection collection.



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓	📈	✎ 🗑

Figure 8-28 Collections view showing the editing and monitoring options

5. Click the **Crawler** tab.
6. Click **Create crawler** and create a crawler to a content repository that contains at least 1000 documents.

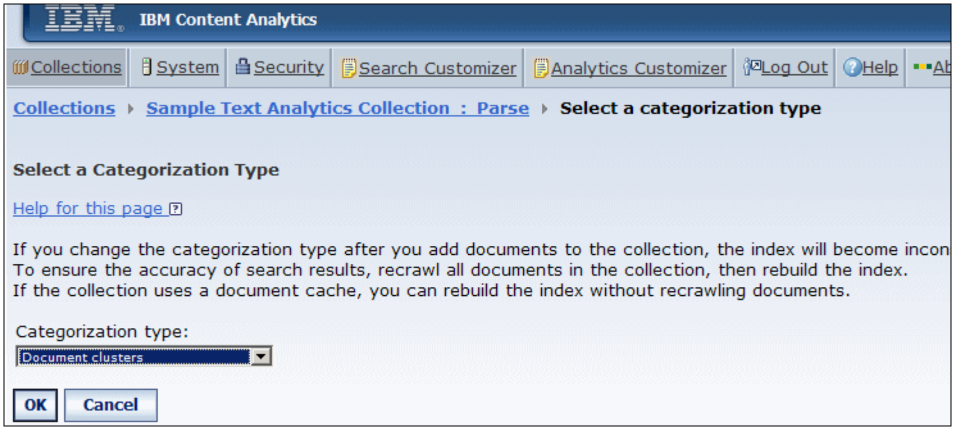
   > **Index for document clustering:** Document clustering requires 1,000 to less than 10,000 documents in the index. You must wait until the index has completed before creating the cluster proposal.

7. Click the **Parse and Index** tab.

8. Click **Select a categorization type** to see the categorization type for the collection that will be displayed.
9. In the Select a Categorization Type pane, under Categorization type, select another value from the Categorization type field and click **OK** (Figure 8-29).

   If you change the categorization type from the **None** or **Rule based** options to **Document clusters** or **Rule based and Document clusters**, rebuild the index. Click **Restart a full index build**.

   However, because we already enabled document clustering when we created the collection, click **Cancel** to return to the Parse and Index edit pane.



Select a Categorization Type

If you change the categorization type after you add documents to the collection, the index will become incon
To ensure the accuracy of search results, recrawl all documents in the collection, then rebuild the index.
If the collection uses a document cache, you can rebuild the index without recrawling documents.

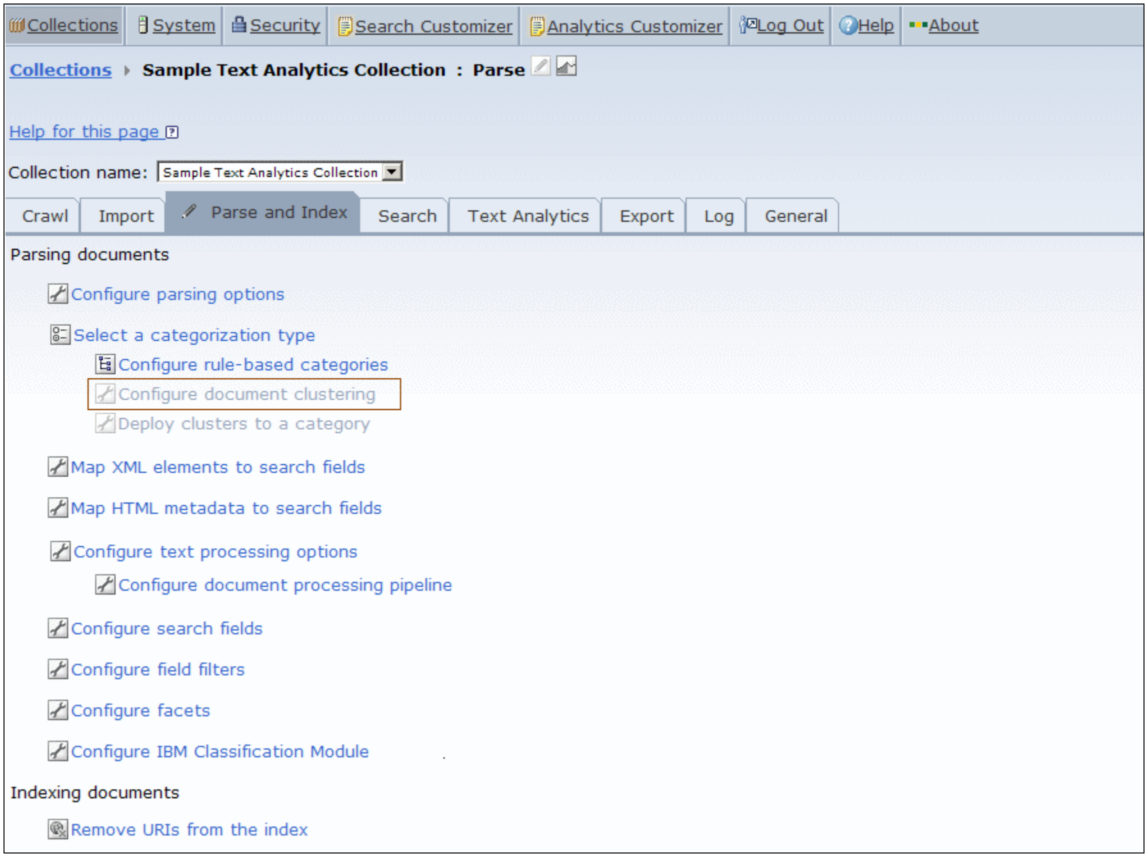
Categorization type: Document clusters

Figure 8-29 Enabling the categorization type

8.3.2 Creating a cluster proposal

To create a cluster proposal, follow these steps:

1. In the Parse and Index pane (Figure 8-30), click **Configure document clustering** to configure the document clustering settings.



Collections | System | Security | Search Customizer | Analytics Customizer | Log Out | Help | About

Collections ▸ **Sample Text Analytics Collection : Parse**

Help for this page

Collection name: Sample Text Analytics Collection

Crawl | Import | Parse and Index | Search | Text Analytics | Export | Log | General

Parsing documents

- Configure parsing options
- Select a categorization type
  - Configure rule-based categories
  - Configure document clustering
  - Deploy clusters to a category
- Map XML elements to search fields
- Map HTML metadata to search fields
- Configure text processing options
  - Configure document processing pipeline
- Configure search fields
- Configure field filters
- Configure facets
- Configure IBM Classification Module

Indexing documents

- Remove URIs from the index

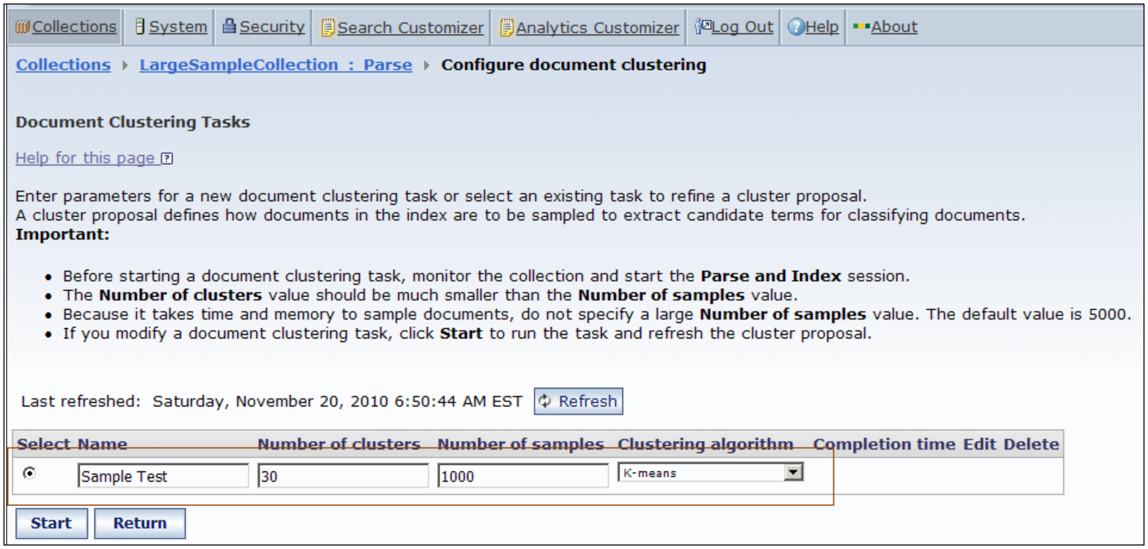
Figure 8-30 Configure document clustering option

2. In the Document Clustering Tasks pane (Figure 8-31), complete the following steps:

   a. In the Name field, type `Sample Test`
   b. In the Number of clusters field, type `30`.
   c. In the Number of samples field, type `1000`.

   > **Number of samples:** The number of samples value must be less than the number of documents in the index. It must be a number from 1,000 to 10,000.

   d. In the Clustering algorithm field, select **K-means**.
   e. Click **Start**.



Collections ▸ LargeSampleCollection : Parse ▸ **Configure document clustering**

Document Clustering Tasks

Help for this page

Enter parameters for a new document clustering task or select an existing task to refine a cluster proposal.
A cluster proposal defines how documents in the index are to be sampled to extract candidate terms for classifying documents.
Important:

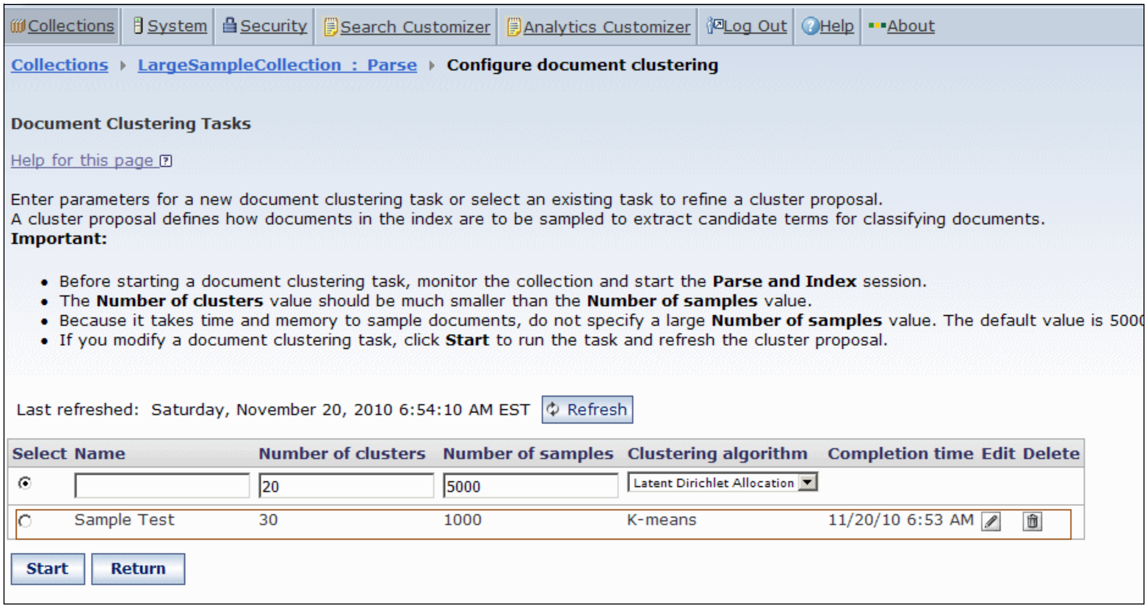
- Before starting a document clustering task, monitor the collection and start the **Parse and Index** session.
- The **Number of clusters** value should be much smaller than the **Number of samples** value.
- Because it takes time and memory to sample documents, do not specify a large **Number of samples** value. The default value is 5000.
- If you modify a document clustering task, click **Start** to run the task and refresh the cluster proposal.

Last refreshed: Saturday, November 20, 2010 6:50:44 AM EST Refresh

Select	Name	Number of clusters	Number of samples	Clustering algorithm	Completion time	Edit	Delete
◉	Sample Test	30	1000	K-means			

Figure 8-31 Setting the parameters in the Document Clustering Tasks pane

3. Monitor the clustering process by clicking **Refresh** until the process is finished (Figure 8-32).



Collections ▸ LargeSampleCollection : Parse ▸ **Configure document clustering**

Document Clustering Tasks

Help for this page

Enter parameters for a new document clustering task or select an existing task to refine a cluster proposal.
A cluster proposal defines how documents in the index are to be sampled to extract candidate terms for classifying documents.
Important:

- Before starting a document clustering task, monitor the collection and start the **Parse and Index** session.
- The **Number of clusters** value should be much smaller than the **Number of samples** value.
- Because it takes time and memory to sample documents, do not specify a large **Number of samples** value. The default value is 500
- If you modify a document clustering task, click **Start** to run the task and refresh the cluster proposal.

Last refreshed: Saturday, November 20, 2010 6:54:10 AM EST

Select	Name	Number of clusters	Number of samples	Clustering algorithm	Completion time	Edit	Delete
●		20	5000	Latent Dirichlet Allocation			
○	Sample Test	30	1000	K-means	11/20/10 6:53 AM	✎	🗑

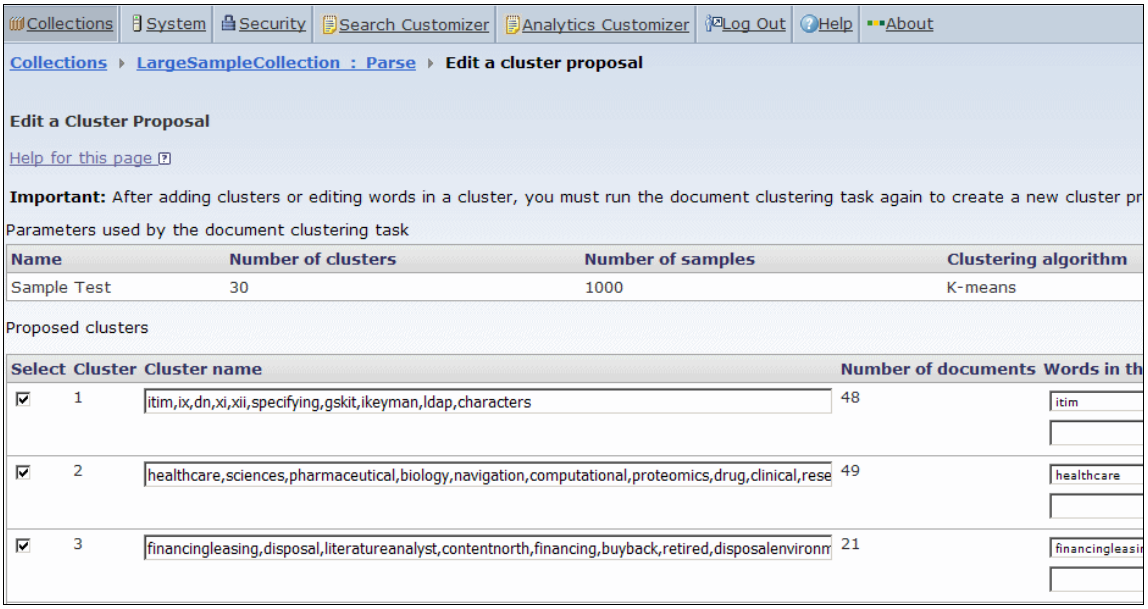
Figure 8-32 Monitoring the document clustering process

8.3.3 Refining the cluster results

After inspecting the initial document clustering results by reviewing the proposed clusters, you can further refine the results:

1. When the cluster task is complete, click the **Edit** icon associated with the Sample Test cluster task to edit a cluster proposal.

The Edit a Cluster Proposal pane (Figure 8-33) is now displayed. The cluster name usually refers to the most popular typical terms in the cluster. The names give you a general idea about the cluster content.



Edit a Cluster Proposal

Help for this page

Important: After adding clusters or editing words in a cluster, you must run the document clustering task again to create a new cluster pr

Parameters used by the document clustering task

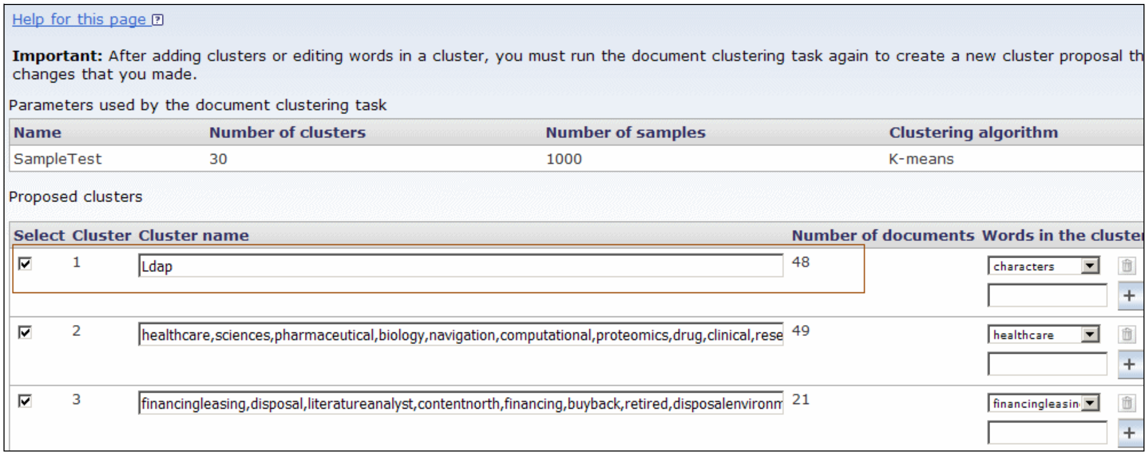
Name	Number of clusters	Number of samples	Clustering algorithm
Sample Test	30	1000	K-means

Proposed clusters

Select	Cluster	Cluster name	Number of documents	Words in th
☑	1	itim,ix,dn,xi,xii,specifying,gskit,ikeyman,ldap,characters	48	itim
☑	2	healthcare,sciences,pharmaceutical,biology,navigation,computational,proteomics,drug,clinical,rese	49	healthcare
☑	3	financingleasing,disposal,literatureanalyst,contentnorth,financing,buyback,retired,disposalenvironm	21	financingleasir

Figure 8-33 Reviewing and editing a clustering proposal result

2. To rename a document cluster, type the new name in the cluster name field, as shown in Figure 8-34.



Help for this page

Important: After adding clusters or editing words in a cluster, you must run the document clustering task again to create a new cluster proposal th changes that you made.

Parameters used by the document clustering task

Name	Number of clusters	Number of samples	Clustering algorithm
SampleTest	30	1000	K-means

Proposed clusters

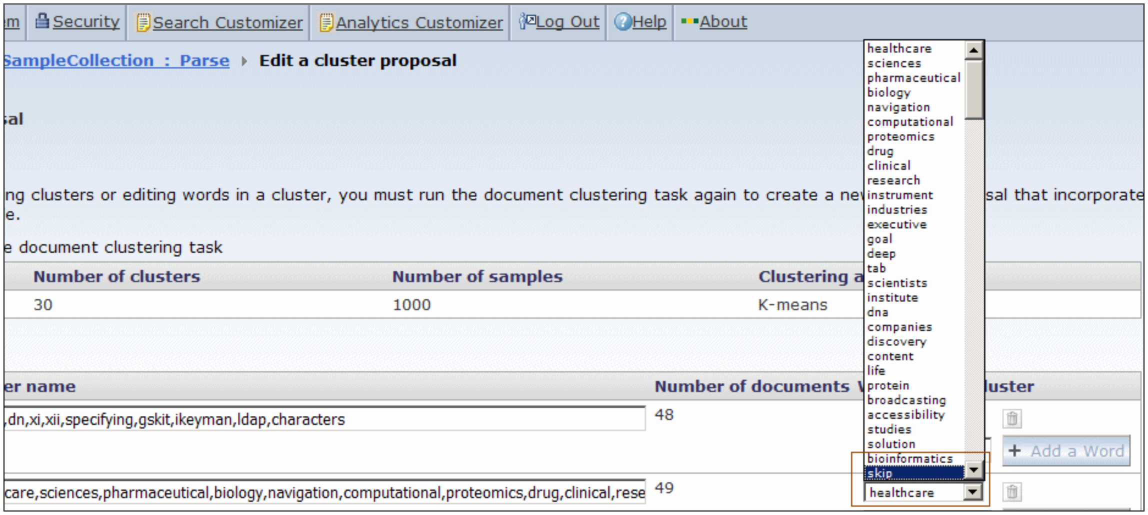
Select	Cluster	Cluster name	Number of documents	Words in the cluste
☑	1	Ldap	48	characters
☑	2	healthcare,sciences,pharmaceutical,biology,navigation,computational,proteomics,drug,clinical,rese	49	healthcare
☑	3	financingleasing,disposal,literatureanalyst,contentnorth,financing,buyback,retired,disposalenvironm	21	financingleasin

Figure 8-34 Renaming clusters

The words listed under “Words in the cluster” contain the typical terms in the cluster and are ranked in popular order. The terms that are most popular are listed first.

3. To remove any cluster group, clear the check box in the **Select** column that is associated with the cluster you want to remove.
4. After you complete the changes for the cluster names and delete any unnecessary cluster groups, click **OK**.
5. In the Document Clustering Tasks pane (Figure 8-32 on page 348), select the **Sample Text** task. Then click **Start** to start the document clustering task.

   Alternatively, you can click **Cancel**. In this case, no other clustering task is run. The changes you made become effective after you deploy the clustering proposal, as explained in step 8 on page 345.
6. Click **Refresh** until the process displays are complete (Figure 8-32 on page 348).
7. Click the **Edit** icon associated with the Sample Test cluster task to edit a cluster proposal (Figure 8-32 on page 348).
8. In the Edit a cluster proposal pane (Figure 8-35), add a word to the document cluster.
   a. In the Words in cluster column, select the new word in the text field, and click **Add a Word**.



SampleCollection : Parse ▸ Edit a cluster proposal

al

ng clusters or editing words in a cluster, you must run the document clustering task again to create a ne sal that incorporate e.

e document clustering task

Number of clusters	Number of samples	Clustering a
30	1000	K-means

er name	Number of documents	luster
,dn,xi,xii,specifying,gskit,ikeyman,ldap,characters	48	
care,sciences,pharmaceutical,biology,navigation,computational,proteomics,drug,clinical,rese	49	

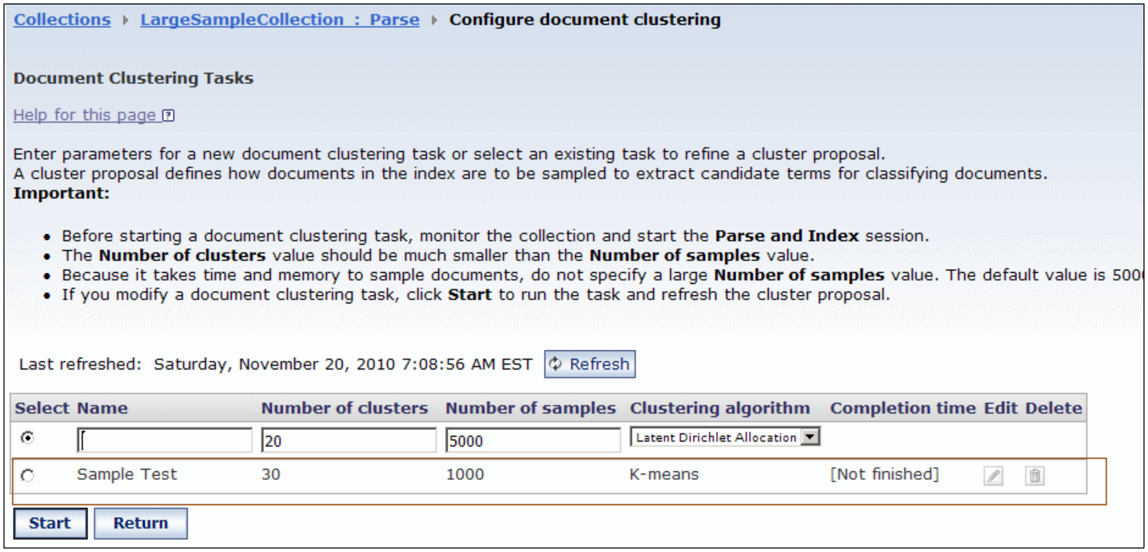
Figure 8-35 Refine the document clustering results by choosing new words and expressions

b. To delete defined words in the cluster, under Words in cluster, select the word from the list. Click the **Delete** icon associated to that particular cluster. As a result, the cluster is refined to better represent your use case.

c. Optional: Add a cluster by clicking **Add a Cluster**.

d. After you make the desired changes to the cluster proposal, click **OK**.

9. Rerun the document clustering process by selecting the cluster name and clicking **Start** (Figure 8-36). Click **Refresh** to monitor the clustering run.



Collections ▸ LargeSampleCollection : Parse ▸ **Configure document clustering**

Document Clustering Tasks

Help for this page

Enter parameters for a new document clustering task or select an existing task to refine a cluster proposal.
A cluster proposal defines how documents in the index are to be sampled to extract candidate terms for classifying documents.
Important:

- Before starting a document clustering task, monitor the collection and start the **Parse and Index** session.
- The **Number of clusters** value should be much smaller than the **Number of samples** value.
- Because it takes time and memory to sample documents, do not specify a large **Number of samples** value. The default value is 500
- If you modify a document clustering task, click **Start** to run the task and refresh the cluster proposal.

Last refreshed: Saturday, November 20, 2010 7:08:56 AM EST

Select	Name	Number of clusters	Number of samples	Clustering algorithm	Completion time	Edit	Delete
●		20	5000	Latent Dirichlet Allocation			
○	Sample Test	30	1000	K-means	[Not finished]		

Figure 8-36 Monitoring the document clustering run

The process of refining the document cluster might take more than one iteration. Continue to refine the clusters until you are satisfied with the result.

8.3.4 Deploying clusters to a category

Document categorization that is based on document clustering involves the following tasks:

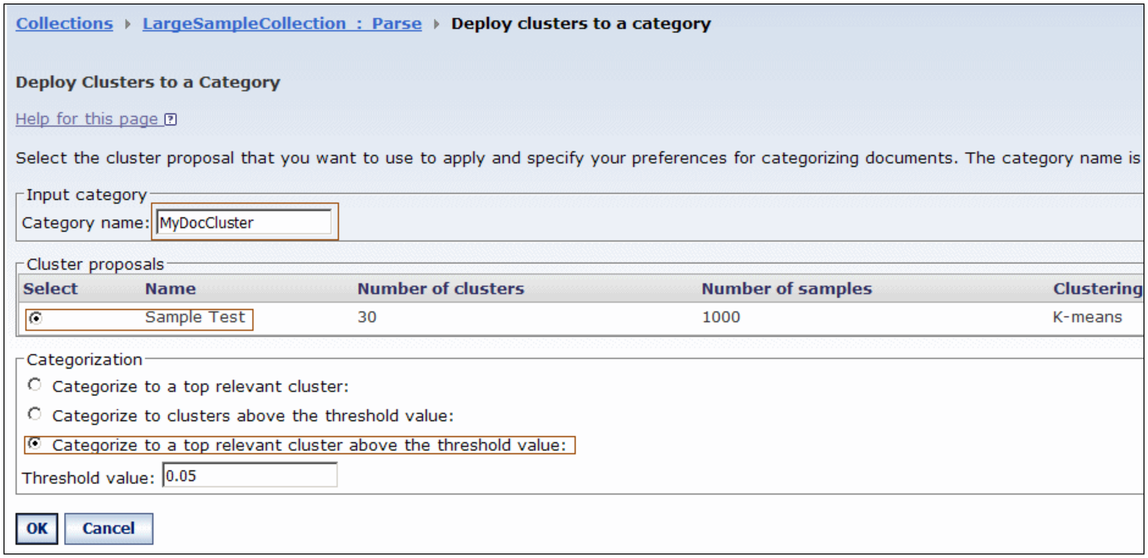
- Configuring the system to detect clusters by sampling a subset of documents and extracting words. See 8.3.1, “Setting up document cluster” on page 343, and 8.3.3, “Refining the cluster results” on page 348.

> **Renaming cluster names:** Before you annotate the entire collection, decide on the most appropriate cluster names. Rename them by double-clicking their names and typing a new value.

- Deploying a document categorization task to add metadata to documents based on the cluster analysis. In this process, the internal Classification Module knowledge base is created. The knowledge base is used to classify all documents in the index into rule-based categories.

To deploy a document categorization task (annotate a full collection), follow these steps:

1. Click the **Parse and Index** tab.
2. If you are not in edit mode, click the **Edit** icon.
3. Click **Deploy clusters to a category**.
4. In the Deploy clusters to a category pane (Figure 8-37), enter the following information:
   a. In the Category name field, type `MyDocCluster`.
   b. Select the cluster set that you want deployed. For our scenario, we select **Sample Test**.
   c. Select a categorization type. For our scenario, we select **Categorize to a top relevant cluster above the threshold value**.
   d. Click **OK**.



Collections ▸ LargeSampleCollection : Parse ▸ Deploy clusters to a category

Deploy Clusters to a Category

Help for this page

Select the cluster proposal that you want to use to apply and specify your preferences for categorizing documents. The category name is

Input category

Category name: MyDocCluster

Cluster proposals

Select	Name	Number of clusters	Number of samples	Clustering
●	Sample Test	30	1000	K-means

Categorization

○ Categorize to a top relevant cluster:

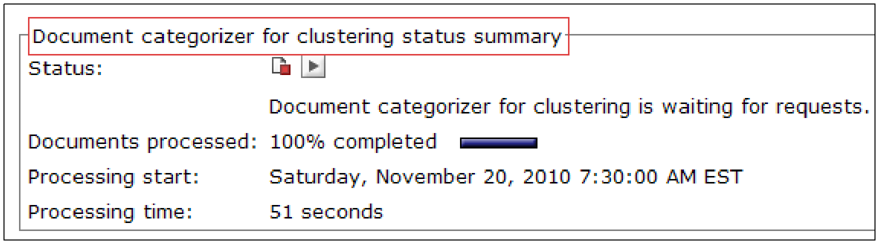
○ Categorize to clusters above the threshold value:

● Categorize to a top relevant cluster above the threshold value:

Threshold value: 0.05

Figure 8-37 Deploying the categorization task

5. Restart the document categorizer:
   a. After this process is finalized, click the **Parse and Index** tab.
   b. Click the **Monitor** icon.
   c. Click **Details**.
   d. In the Document categorizer for clustering status summary section (Figure 8-38), click **Start**. Wait until the document categorizer for clustering process is complete.



Document categorizer for clustering status summary

Status: Document categorizer for clustering is waiting for requests.

Documents processed: 100% completed

Processing start: Saturday, November 20, 2010 7:30:00 AM EST

Processing time: 51 seconds

Figure 8-38 Document categorizer for clustering status summary

6. Deploy the resources to update a category label by clicking **Start** in the Resource deployment status section. Wait until the process is complete.

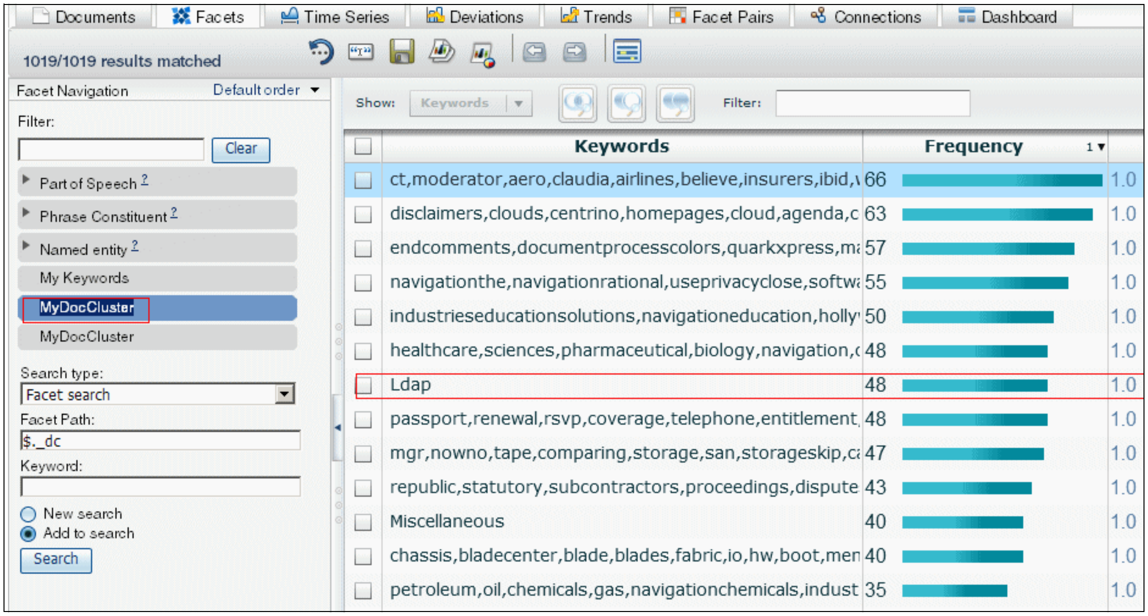
> **Rebuilding the index:** It is not necessary to rebuild the full index. However, you can choose to start an index rebuild for another reason by clicking **Restart a full index build.**

8.3.5 Working with the cluster results

You can now work with the cluster in the search and text miner applications. The deployed cluster analysis can help narrow down your content to help you gain insight by working with the categorized documents.

1. Open the text miner application.
2. Click the **Facets** tab.
3. Click the **Document Cluster** facet. Document Cluster is the default name for the cluster facet.

4. Click the MyDocCluster value that you want to view further. For our scenario, we select **LDAP** (Figure 8-39).



Search type: Facet search

Facet Path: $._dc

Show: Keywords

□	Keywords	Frequency		
□	ct,moderator,aero,claudia,airlines,believe,insurers,ibid,\	66		1.0
□	disclaimers,clouds,centrino,homepages,cloud,agenda,c	63		1.0
□	endcomments,documentprocesscolors,quarkxpress,ma	57		1.0
□	navigationthe,navigationrational,useprivacyclose,softw	55		1.0
□	industrieseducationsolutions,navigationeducation,holly	50		1.0
□	healthcare,sciences,pharmaceutical,biology,navigation,(	48		1.0
□	Ldap	48		1.0
□	passport,renewal,rsvp,coverage,telephone,entitlement,	48		1.0
□	mgr,nowno,tape,comparing,storage,san,storageskip,ca	47		1.0
□	republic,statutory,subcontractors,proceedings,dispute	43		1.0
□	Miscellaneous	40		1.0
□	chassis,bladecenter,blade,blades,fabric,io,hw,boot,men	40		1.0
□	petroleum,oil,chemicals,gas,navigationchemicals,indust	35		1.0

Figure 8-39 Clustering facet on the text miner application

5. Click **Add to search with Boolean AND** to add the LDAP facet to the search query. Now the documents in the result set are limited to those documents within the LDAP cluster group.
6. Click the **Documents** tab to view these documents further.

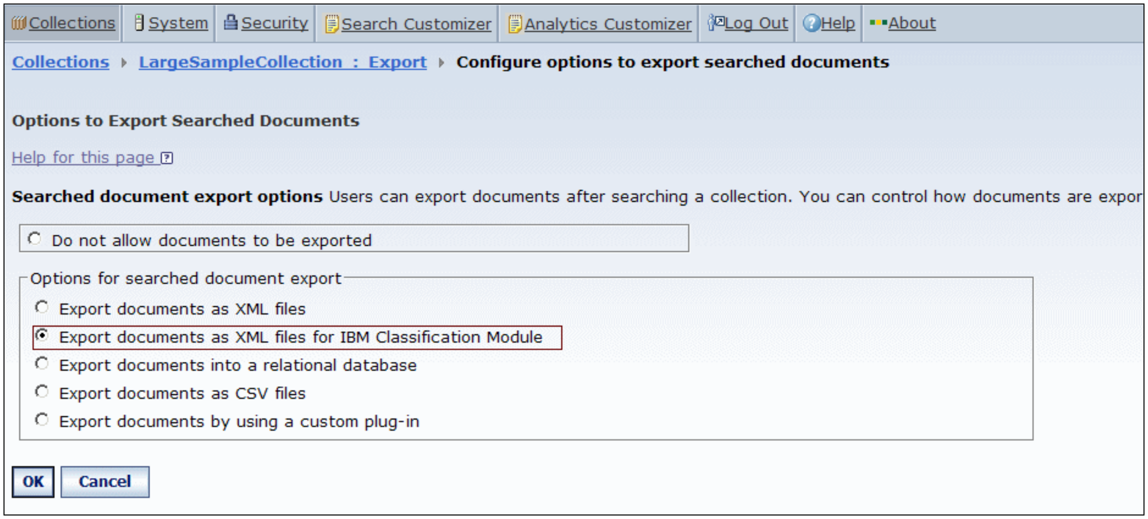
In addition, if a conceptual search is enabled for the collection, you can search documents that conceptually match the query terms. See 9.2.3, “Using a conceptual search for advanced content discovery” on page 364.

8.3.6 Creating and deploying the clustering resource

When you decide that the document cluster generated the expected results, you can use them in Classification Module with the annotator. You can build the Classification Module resource and enable the Classification Module annotator so that future documents are annotated with it.

The document clustering results are stored and presented to the user as facets. The default name of the clustering facet is Document Cluster. When you deploy the categorization task, you have the option of choosing a new name for this dedicated facet. In our scenario, we named the document cluster facet MyDocCluster. You can use the information for this facet to build the Classification Module resources:

1. Export documents under the MyDocCluster facet by using the search export that is configured to export documents as an XML file for Classification Module. This exported data can be used to train the knowledge base. Figure 8-40 shows the search export configured to export as XML files for Classification Module.



Configure options to export searched documents

Options to Export Searched Documents

Help for this page

Searched document export options Users can export documents after searching a collection. You can control how documents are expor

○ Do not allow documents to be exported

Options for searched document export

○ Export documents as XML files
● Export documents as XML files for IBM Classification Module
○ Export documents into a relational database
○ Export documents as CSV files
○ Export documents by using a custom plug-in

Figure 8-40 Exporting XML files for Classification Module

For further information about exporting to Classification Module, see 10.6.3, “Exporting search result documents to the file system for Classification Module” on page 420.

2. Import of the exported XML files into Classification Module Workbench.
3. Create and tune a knowledge base.
4. Build a decision plan to use the document clustering results based on information in the knowledge base and any rules that you create.
5. Export the decision plan to the Classification Module.
6. Enable the Classification Module annotator in Content Analytics.
7. In Content Analytics, map the fields. See 9.3.5, “Configuring the Classification Module annotator” on page 378.

8.3.7 Preferred practices

This section provides guidelines for working with clustering.

When to use document clustering

You might want to use document clustering when you have a lot of unstructured content and little knowledge about it. Without knowing your content, it might be difficult to create dictionaries or obtain valuable insight with the text miner application. Often you must obtain more insight into the content before being able to use the sophisticated text analysis tools.

Document clustering provides insight to a large set of unstructured content without having to configure Classification Module. Content Analytics offers many tools for text analysis, and Clustering is one of them. You are encouraged to use any of the following tools to gain comprehensive insight of your content:

- Leverage the Classification Module annotator built on the top of the clustering based knowledge base, after tuning it by using Classification Module Workbench.
- Build dictionaries using Content Analytics tools.
- Build pattern rules in Content Analytics.
- Generate terms of interest.
- Refine the search results by clustering facets

Number of documents to use in clustering

The granularity of document clustering can influence the nature and quality of the insight. To take advantage of the fullest potential of document clustering insight, follow these steps:

1. Run two or three clusters proposals with different granularity.
2. Inspect the cluster names and refine the results as needed.
3. Inspect the documents represented by a cluster in the text miner application.
4. Decide the best clusters groups to use.
5. Deploy the categorization based on clustering to annotate the entire collection.
6. Use the text miner tools to continue.

Document clustering offers insight about unstructured content that is further used with other text analytics tools. The suggested cluster categories and names offer knowledge to further build dictionaries or patterns rules.

9

Content analysis with IBM Classification Module

IBM Classification Module can classify document content into categories. The Classification Module annotator that is available through IBM Content Analytics enables automatic text classification and context-based text-understanding. This chapter provides details about the Classification Module annotator and explains how to use the annotator for content analysis.

This chapter includes the following sections:

- The Classification Module annotator
- Fine-tuning your analysis with the Classification Module annotator
- Creating and deploying the Classification Module resource
- Validation and maintenance of the Classification Module annotator
- Preferred practices

> **InfoSphere reference:** IBM Classification Module was previously known as *IBM InfoSphere Classification Module*. In this book, you might notice the previous name in some of the application windows and references to the information center for IBM Classification Module.

9.1 The Classification Module annotator

The Classification Module annotator is integrated inside the Unstructured Information Management Architecture (UIMA) document processing pipeline of IBM Content Analytics. The Classification Module annotator uses the capabilities of Classification Module to classify content into categories and generate metadata information that can be used for facets or keywords in Content Analytics.

Classification Module uses sophisticated natural language processing and semantic analysis algorithms to determine the true intent of words and phrases. It then uses that knowledge to automate classification.

Accuracy improves over time because the system adapts to your content by identifying different categories from examples that you provide. When you provide feedback, the system adjusts in real time and immediately incorporates any corrections that you make. The accuracy of the classification results keeps pace with changes in your content and environment.

The Classification Module annotator combines contextual statistical analysis with a rule-based, decision-making approach. For example, the system can identify keywords, patterns, and words within a certain proximity of each other. When content that matches a condition in a rule is detected, the action defined for the rule is applied, and content is classified accordingly.

In addition to organizing information by policies or keywords, the Classification Module annotator can also assign metadata that is based on the full context of the document. The classification process searches for a single word or phrase. It also analyzes the entire document, distills the main point of the text, and assigns the text to a category.

9.1.1 When to use the Classification Module annotator

The classification capability of the Classification Module annotator is used to categorize text or make correlations between text and objects (for example, personalization or general data classification applications). Search, text mining, and classification are often integrated together in a single system. They are synergistic for several reasons.

Search, text mining, and classification provide complementary mechanisms for describing documents. *Search* and *text mining* find and describe documents based on a small set of words supplied by users (such as the query “energy bill”). *Classification* attempts to describe the overall document based on a set of descriptors supplied by the taxonomy (for example, one of the subjects in a subject taxonomy). That is, if a search engine supplies the category to the user, it can be easy for the user to distinguish which search results are relevant.

For example, if the user query is “energy bill,” some of the results are marked as a piece of energy legislation, but others are marked as a monthly electric or gas bill. A user seeing this mixture of topics can then refine the query to select just the ones intended by this ambiguous query.

Search and classification can be paired in the following ways:

- Search within a category. You can select a category and then search only documents that are both within the category and that match your query.
- Facet search. In this method, you can specify several different facets (or characteristics) of a document to a search engine (for example, “search for all PDF documents about databases from last year”). This search is a generalization of “search within a category,” where multiple criteria that might or might not be categories from a taxonomy can be combined.
- Taxonomy browsing. Some or all of the documents on a website are displayed as a taxonomy that can be navigated, with each document assigned to one or more nodes of the taxonomy.
- Classifying search results. The results of a search are displayed together with their assigned categories. Categories can be used to group or sort result sets.

9.1.2 The Classification Module technology

Understanding and classifying text is an old problem in the field of artificial intelligence. Determining the most likely category in which to classify a new content item is not trivial.

With IBM classification technology, applications can understand and classify unstructured free-form text. The Classification Module annotator attempts to understand information based on its existing knowledge. It “learns” how to distinguish and classify data based on its acquired information. For example, the technology learns how to distinguish between text about dogs and cats, after you provide it with example text about dogs and another set of example text about cats. Then, the system attempts to correctly classify the new text as being related to dogs or cats.

The Classification Module annotator learns from real-world examples and stores classification information into what is referred to as a *knowledge base*. The Classification Module annotator consults the knowledge base to classify new text into categories based on their similarity to the text seen in the past.

To create the knowledge base, the Classification Module annotator is first trained by using a body of sample data, such as emails, documents or other text, that has been preclassified into appropriate categories. This body of data is known as a *corpus*. The corpus consists of sample text that represents the kind of information that the system is expected to classify. It creates statistical models that make up the knowledge base of a system.

After the Classification Module annotator is trained, new text can be submitted for classification by using a process called *matching*. The Classification Module annotator analyzes the new text and computes relevancy scores for each category in the knowledge base, as a measure of how closely the text matches each category.

After creating and training a knowledge base, you can build a Classification Module decision plan that contains rules that refer to the knowledge base suggestions. The decision plan can also extract metadata that is associated with the text.

Figure 9-1 shows an example of importing a sample set of documents to the Classification Module annotator to create and train a knowledge base to recognize different types of burn documents. The types of burns include corneal burns, fire burns, and chemical burns.

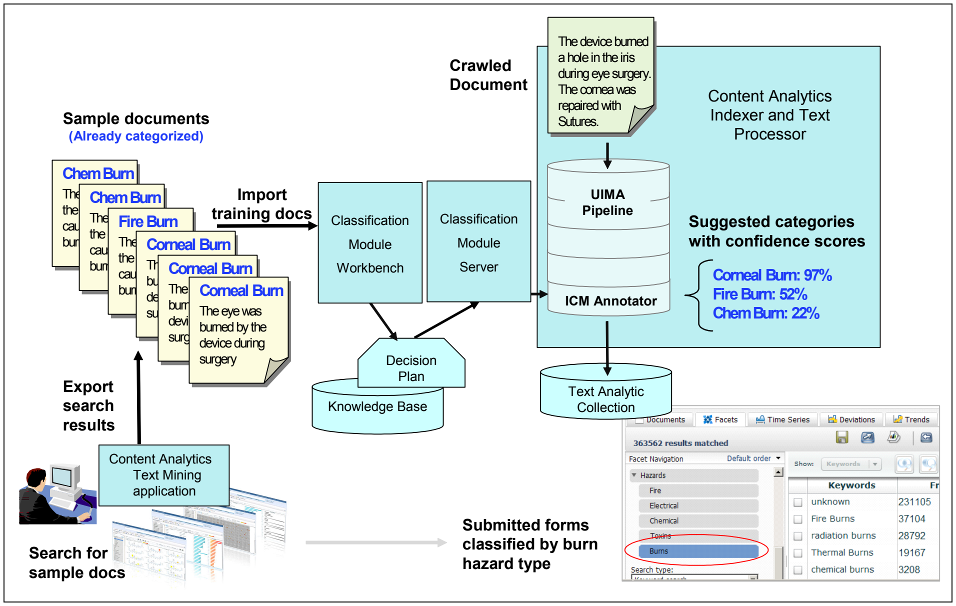


Figure 9-1 Integration of Classification Module and Content Analytics

After the knowledge base is trained, you use the Classification Module annotator to categorize all the documents in terms of various burns with a metadata set on the documents. The metadata information is then used for Hazard type → Burn facet (which is also called the *Burn Hazard facet* for simplicity). You can then use Content Analytics to analyze the different types of burns and any correlation to other factors and discover important insights related to the burns or other hazard-related matters.

9.2 Fine-tuning your analysis with the Classification Module annotator

This section explains how to use the Classification Module annotator to fine-tune your content analysis.

9.2.1 Building your collection

The first step when working with the Classification Module annotator is to gather sample content and categorize the content into buckets that represent different categories. The content and the associated categories are then used to train a Classification Module knowledge base.

The example uses a set of documents that contain information related to a Fictitious Medical Devices Company A. The data set describes various adverse events for medical devices. The company wants to gain insight into its content and wants to quickly identify quality control issues and fix them. The company also wants to be prepared for legal and regulatory actions.

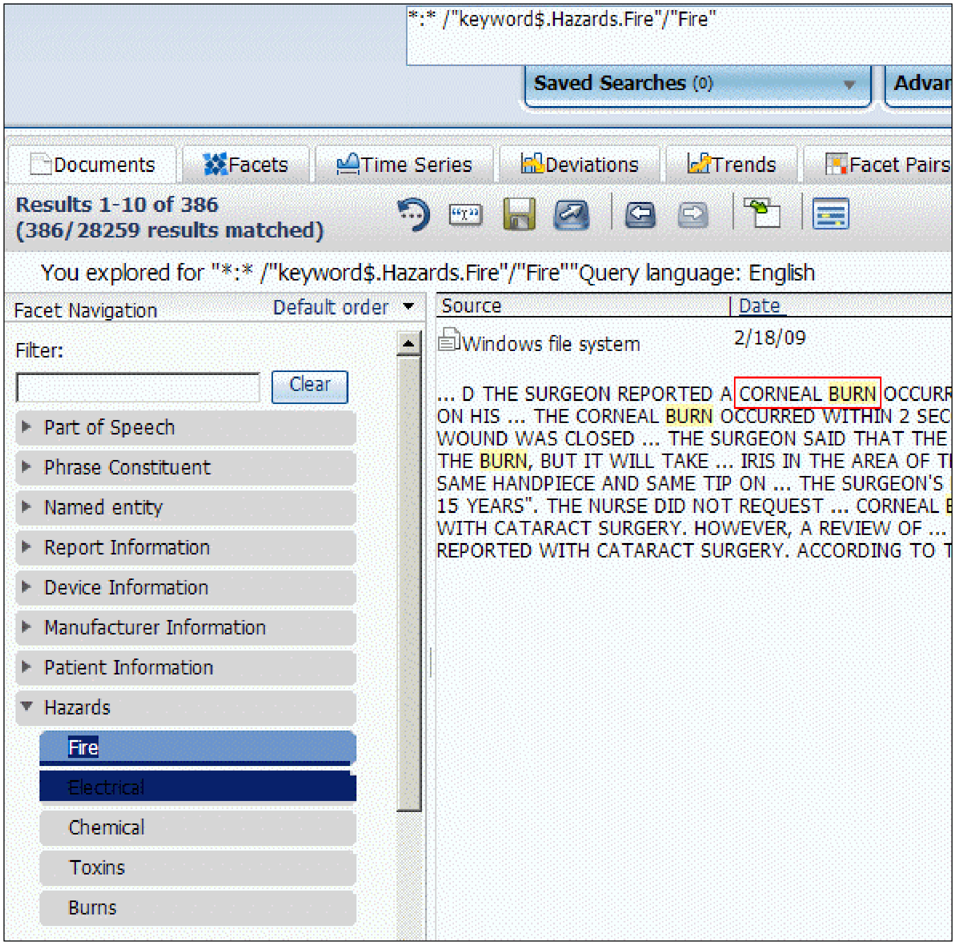
To analyze the content, follow the standard procedure for building a text analytics collection by using the documents of the manufacturer:

1. Build a collection from the documents. The documents come with a list of fields with well-defined values, such as equipment name and type. They also come with a series of textual fields that contain free format text information.
2. Define facets and associate them with data fields and textual fields from the content.
3. Enable the **Named Entity Recognition** annotator.
4. Define new facets and associate them with dictionary entries that are relevant to the content case of the manufacturer. For this example, you define a new facet called *Hazards* that categorizes all hazards caused by the failing devices: fire, chemical, toxic, or electrical. You also create synonyms to be

used in the dictionary. For example, for fire-related hazards, you create the synonyms “burn,” “smoke,” “blaze,” and “flame” for “fire.”

5. Use the text miner application of Content Analytics to find problems related to the medical devices and procedures described in the documents.

By looking at the documents under the Fire Hazards facet, you see several documents that do not refer specifically to hazards caused by fire. However, these documents are in the view because the word “burn” was found in them, and they were labeled as such. Finding documents that have the words, but where the usage is unrelated, is one of the drawbacks of the rule-based dictionary approach to classification. In this example, you find a series of documents related to “corneal burn,” similar to the example in Figure 9-2, that do not have anything to do with a fire hazard in these scenarios.



: /"keyword$.Hazards.Fire"/"Fire"

Results 1-10 of 386
(386/28259 results matched)

You explored for "*:* /"keyword$.Hazards.Fire"/"Fire""Query language: English

Source	Date
Windows file system	2/18/09

... D THE SURGEON REPORTED A CORNEAL BURN OCCURR
ON HIS ... THE CORNEAL BURN OCCURRED WITHIN 2 SEC
WOUND WAS CLOSED ... THE SURGEON SAID THAT THE
THE BURN, BUT IT WILL TAKE ... IRIS IN THE AREA OF T
SAME HANDPIECE AND SAME TIP ON ... THE SURGEON'S
15 YEARS". THE NURSE DID NOT REQUEST ... CORNEAL B
WITH CATARACT SURGERY. HOWEVER, A REVIEW OF ...
REPORTED WITH CATARACT SURGERY. ACCORDING TO T

Figure 9-2 Example of an unrelated search result in the Fire Hazards facet view

9.2.2 Refining the analysis

Because you encountered several documents unrelated to the Fire Hazard facet, you must refine the analysis and look for more accurate techniques for text mining. The Classification Module annotator can help you to improve the overall accuracy of classification. Classification Module has the unique ability to consider the entire context of the document and not just a few words that are in the document. With this ability, Classification Module can differentiate better between the different types of burns (chemical burns, radiation burns, and thermal burns). In the example, you use the Classification Module annotator to learn the difference between the different types of burns and to correctly categorize them.

First, you gather a few examples of documents for training the Classification Module annotator to identify these different types of *burn hazards*. For the example, you use the export capabilities of Content Analytics Version 2.1 (as explained in 10.6.3, "Exporting search result documents to the file system for Classification Module" on page 420). You use them to gather relevant content for training the knowledge base and designing the decision plan in Classification Module:

1. Use the text miner application in Content Analytics to identify sample documents for each of the categories: chemical burn, radiation burn, fire burn, and thermal burn.
2. Search for "`chemical burn`" and review the documents.
3. After you are satisfied with the example documents that you find in the text miner application, click **Export**. Make sure to provide a meaningful name for the XML file and Description fields. The name you assign for the Description field becomes the Classification Module Category label for the current set of examples. Therefore, the name must be meaningful and exact.
4. Repeat steps 2 and 3 to gather examples for all the categories that you plan for Classification Module to learn. When you finalize this process, you have a directory with a series of XML files that contain sample documents for training a Classification Module knowledge base. The `catalog.xml` file contains the information regarding the sample data fields names.

To use the Classification Module annotator, create the following resources in Classification Module and publish them to the Classification Module server:

- Knowledge base
- Decision plan related to the knowledge base

9.2.3 Using a conceptual search for advanced content discovery

In a traditional text search, both documents and queries are regarded as sets of terms. A document is a good match for a search query if the document contains terms of that query. Documents with more terms in common with the query are ranked higher in the search result. If you need to discover documents that are conceptually similar, even if the query terms do not exactly match the document content, you can enable a conceptual search within the collection.

For more information about setting up the conceptual search, go to IBM Content Analytics Information Center at the following address, and search on *configuring Classification Module search fields and scores*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

Through the process of a concept-based search, you can discover documents that are the most similar to the query in terms of their classification results. Classification Module is used in Content Analytics to categorize documents and assign them a relevancy score for each category suggested in the results set. Based on this information, the conceptual search returns its results.

In addition, the document categorization can be based on the clustering discovery. The conceptual search can be run against the categorization determined by clustering for the collection to further refine the categorization.

Using a conceptual search in Content Analytics

Conceptual searches can be used in Content Analytics for the following reasons:

- To improve search ranking based on conceptual resemblance. Documents that conceptually resemble a search query are regarded as highly relevant in the search results. This type of search helps to improve the quality of the search results ranking and helps users to find documents they are looking for without knowing the exact terms contained in the documents.
- To filter documents that conceptually match a query regardless of whether these documents contain the search query terms. If documents do not contain any query terms, but contain terms that are used frequently with the query terms, they are included in the search results.

Reference materials: For information about building and maintaining a decision plan and knowledge base in Classification Module, see the following resources:

- IBM InfoSphere Classification Module tutorial

  http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp?topic=/com.ibm.classify.tutorial.doc/bnrtu000.htm

- IBM InfoSphere Classification Module Information Center

  http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp

For hints and tips about building a knowledge base and decision plan, see *IBM Classification Module: Make It Work for You*, SG24-7707.

9.3 Creating and deploying the Classification Module resource

This section explains the step-by-step procedures for creating the knowledge bases and decision plan in Classification Module and for deploying them into the Content Analytics document processing pipeline. It does not attempt to teach you how to do everything in Classification Module. For comprehensive Classification Module product usage, see the materials referenced in the previous shaded box.

Creating and deploying Classification Module resources entails the following tasks, which are explained in the sections that follow:

1. Starting the Classification Module server
2. Creating and training the knowledge bases
3. Creating a decision plan
4. Deploying the knowledge base and decision plan
5. Configuring the Classification Module annotator

9.3.1 Starting the Classification Module server

Before you create the knowledge base and decision plan, start the Classification Module server:

1. Click the **Services** icon in your task bar to open the Services Management Console in Windows.
2. Make sure that the IBM Classification Module Process Manager is running. If it is not running, in Windows, right-click **IBM Classification Module Process Manager** and select **Start the service**.

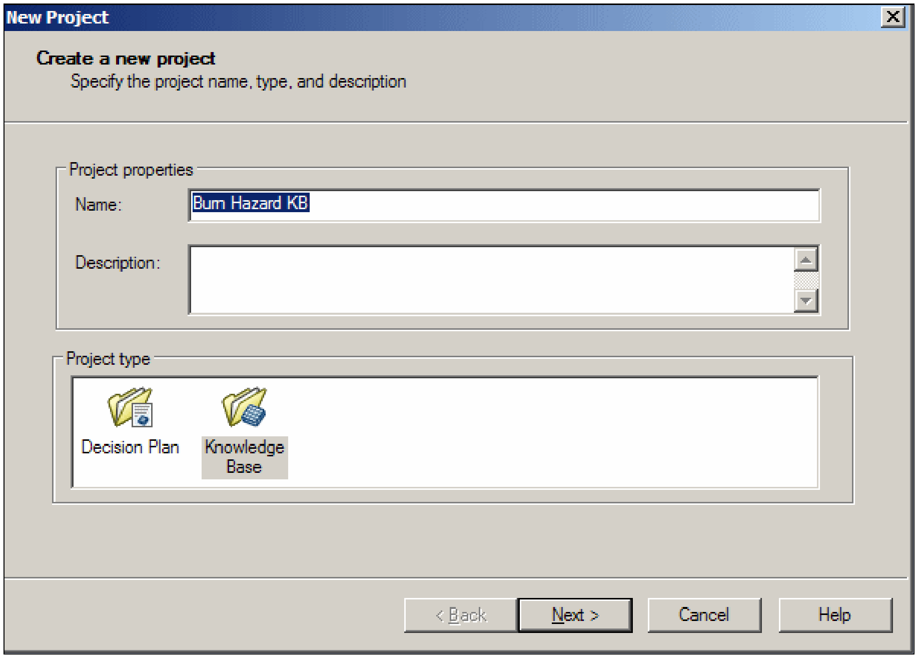
3. Open the Classification Module Management Console by selecting **Start** → **All Programs** → **IBM Classification Module 8.7** → **Management Console**.

The Classification Module Management Console is the Classification Module Server administration tool with which you can manage all the knowledge bases and decision plans. In this chapter, Classification Module Management Console is used to ensure that the decision plans and knowledge bases that are needed for the Content Analytics collections are installed on the server and running.

9.3.2 Creating and training the knowledge bases

To create a knowledge base, follow these steps:

1. Start the Classification Workbench by selecting **Start** → **All Programs** → **IBM Classification Module 8.7** → **Classification Workbench**.
2. Select **Project** → **New** → **Knowledge Base**.
3. In the New Project window (Figure 9-3), type the name of the knowledge base. For this example, enter `Burn Hazard KB`. Then click **Next**.



New Project

Create a new project

Specify the project name, type, and description

Project properties

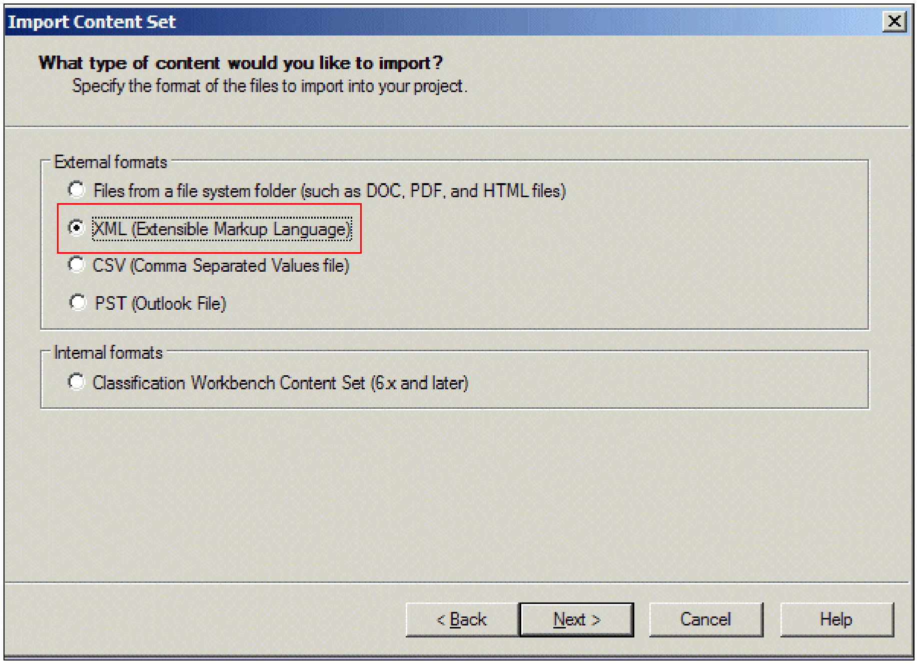
Name: Burn Hazard KB

Description:

Project type: Knowledge Base

Figure 9-3 Creating a Classification Module knowledge base project

4. Import the XML files that were exported by Content Analytics. Follow the defaults of the wizard until you reach the next step to import the content set.
5. In the Import Content Set window (Figure 9-4), select **XML** and click **Next**.



Import Content Set

What type of content would you like to import?

Specify the format of the files to import into your project.

External formats

○ Files from a file system folder (such as DOC, PDF, and HTML files)

● XML (Extensible Markup Language)

○ CSV (Comma Separated Values file)

○ PST (Outlook File)

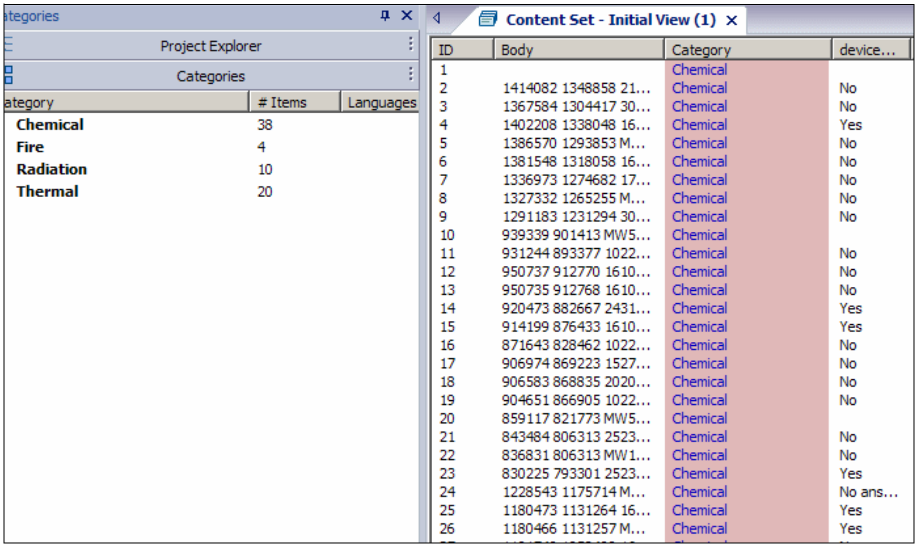
Internal formats

○ Classification Workbench Content Set (6.x and later)

Figure 9-4 Importing XML files that were exported by Content Analytics

6. Select the XML folder where you exported the data from Content Analytics. Because you have the `catalog.xml` file, Classification Module Workbench has all the necessary information regarding the fields. Click **Next**.
7. In the next window, click **Finish**.

Figure 9-5 shows the full content set. The Category field is highlighted in pink.



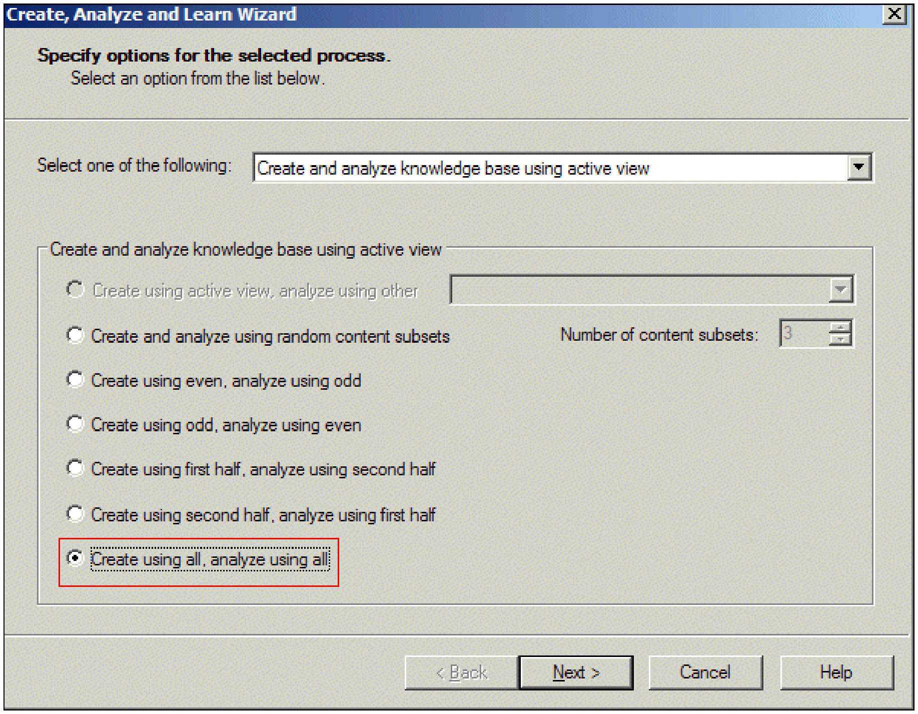
ategory	# Items	Languages
Chemical	38	
Fire	4	
Radiation	10	
Thermal	20	

ID	Body	Category	device...
1		Chemical	
2	1414082 1348858 21...	Chemical	No
3	1367584 1304417 30...	Chemical	No
4	1402208 1338048 16...	Chemical	Yes
5	1386570 1293853 M...	Chemical	No
6	1381548 1318058 16...	Chemical	No
7	1336973 1274682 17...	Chemical	No
8	1327332 1265255 M...	Chemical	No
9	1291183 1231294 30...	Chemical	No
10	939339 901413 MW5...	Chemical	
11	931244 893377 1022...	Chemical	No
12	950737 912770 1610...	Chemical	No
13	950735 912768 1610...	Chemical	No
14	920473 882667 2431...	Chemical	Yes
15	914199 876433 1610...	Chemical	Yes
16	871643 828462 1022...	Chemical	No
17	906974 869223 1527...	Chemical	No
18	906583 868835 2020...	Chemical	No
19	904651 866905 1022...	Chemical	No
20	859117 821773 MW5...	Chemical	
21	843484 806313 2523...	Chemical	No
22	836831 806313 MW1...	Chemical	No
23	830225 793301 2523...	Chemical	Yes
24	1228543 1175714 M...	Chemical	No ans...
25	1180473 1131264 16...	Chemical	Yes
26	1180466 1131257 M...	Chemical	Yes

Figure 9-5 Classification Module Workbench showing the imported sample data

8. Open the Create, Analyze and Learn Wizard to train the new knowledge base.

9. In the Specify options for the selected process window (Figure 9-6), select **Create using all, analyze using all** and click **Next**.



Create, Analyze and Learn Wizard

Specify options for the selected process.
Select an option from the list below.

Select one of the following: Create and analyze knowledge base using active view

Create and analyze knowledge base using active view

○ Create using active view, analyze using other

○ Create and analyze using random content subsets — Number of content subsets: 3

○ Create using even, analyze using odd

○ Create using odd, analyze using even

○ Create using first half, analyze using second half

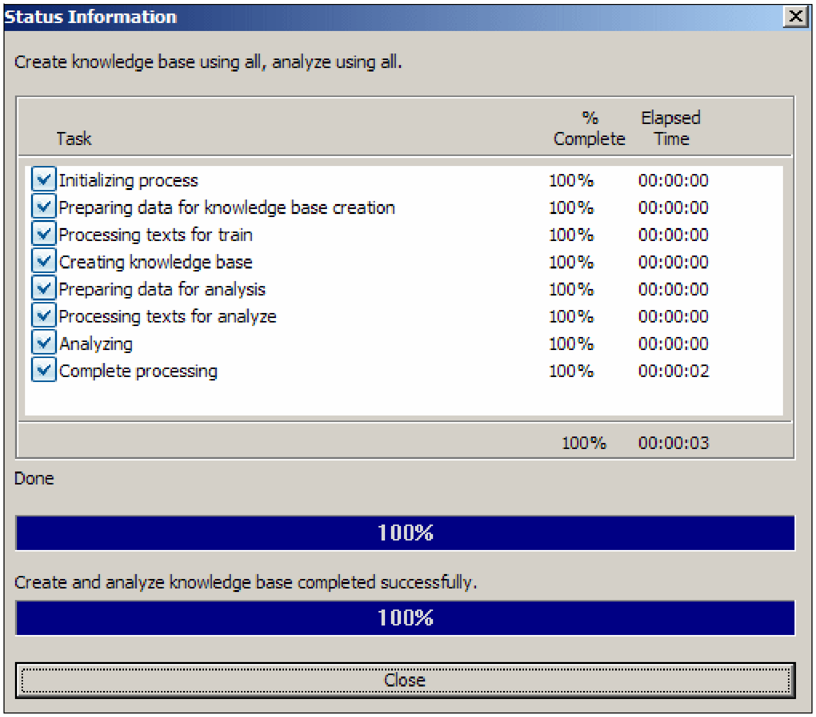
○ Create using second half, analyze using first half

● Create using all, analyze using all

Figure 9-6 Training the Classification Module knowledge base

10. For the next windows, use the default settings until you reach the Status window.

11. In the Status window (Figure 9-7), click **Close**. You have finished creating and training your new knowledge base.



Create knowledge base using all, analyze using all.

	Task	% Complete	Elapsed Time
☑	Initializing process	100%	00:00:00
☑	Preparing data for knowledge base creation	100%	00:00:00
☑	Processing texts for train	100%	00:00:00
☑	Creating knowledge base	100%	00:00:00
☑	Preparing data for analysis	100%	00:00:00
☑	Processing texts for analyze	100%	00:00:00
☑	Analyzing	100%	00:00:00
☑	Complete processing	100%	00:00:02
		100%	00:00:03

Done

100%

Create and analyze knowledge base completed successfully.

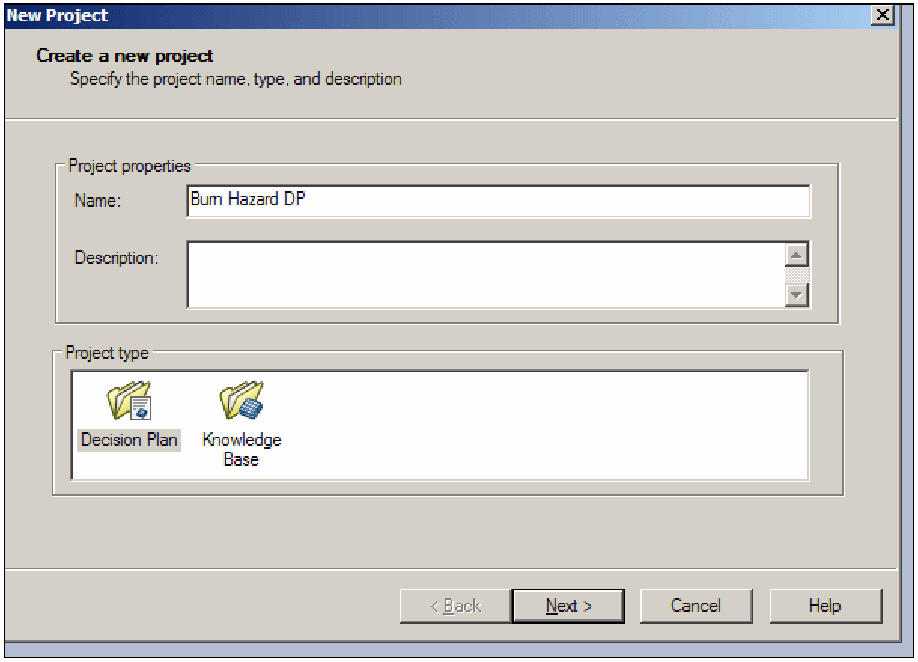
100%

Figure 9-7 Status window showing training is complete

9.3.3 Creating a decision plan

To create a Classification Module decision plan, follow these steps:

1. From Classification Workbench, select **Project** → **New** → **Decision Plan**.
2. In the New Project window (Figure 9-8), type the decision plan name. For this example, enter `Burn Hazard DP`. Then click **Next**.



New Project

Create a new project

Specify the project name, type, and description

Project properties

Name: Bum Hazard DP

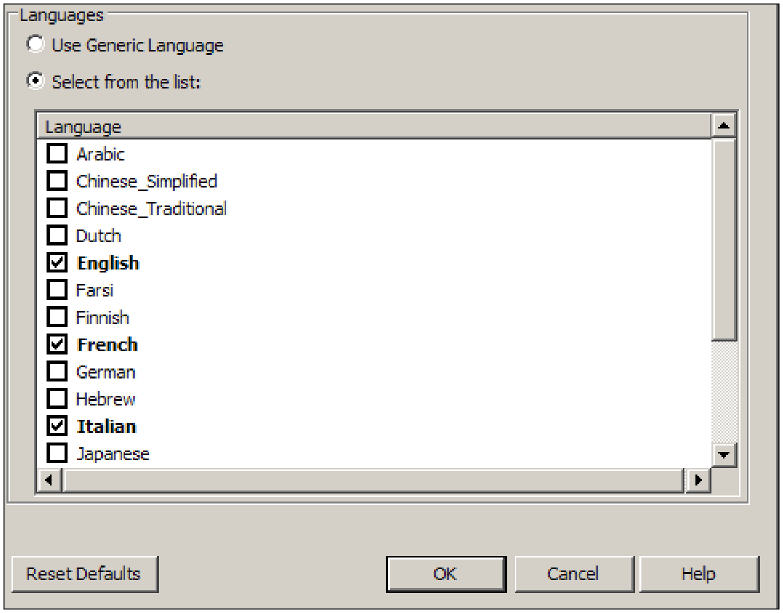
Description:

Project type

Figure 9-8 Creating a Classification Module decision plan project

3. Import the data that was exported by Content Analytics. See 10.6.3, “Exporting search result documents to the file system for Classification Module” on page 420, for details about exporting.
4. When you foresee that your data will be in several languages, go to **Project Explorer**, and select **Project Options**.

5. In the window that opens (Figure 9-9), select all the languages that are relevant for your project. Then click **OK**.



Languages

○ Use Generic Language

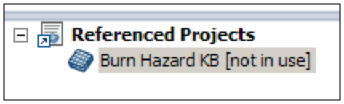
● Select from the list:

Language

- □ Arabic
- □ Chinese_Simplified
- □ Chinese_Traditional
- □ Dutch
- ☑ **English**
- □ Farsi
- □ Finnish
- ☑ **French**
- □ German
- □ Hebrew
- ☑ **Italian**
- □ Japanese

Figure 9-9 Setting the languages for a Workbench project

6. To use the previously created knowledge base, go to **Referenced Projects** and click **Add Project**. Add your knowledge base to the project. For this example, add the **Burn Hazard knowledge base** to the project (Figure 9-10).



⊟ **Referenced Projects**
Burn Hazard KB [not in use]

Figure 9-10 Decision plan under Referenced Projects referring to a knowledge base

7. Define a new rule that sets a new field called *burntype* if the category suggested by the Burn Hazard KB knowledge base has a confidence score above 95:

   a. Right-click **New Group**, and select **New Rule**.

   b. Select **Trigger**. Click **Condition** and choose **Trigger always**.

   c. Select **Actions**. Click **Add** and select **General actions**. Select **Set the value for a content field** and click **Next**.

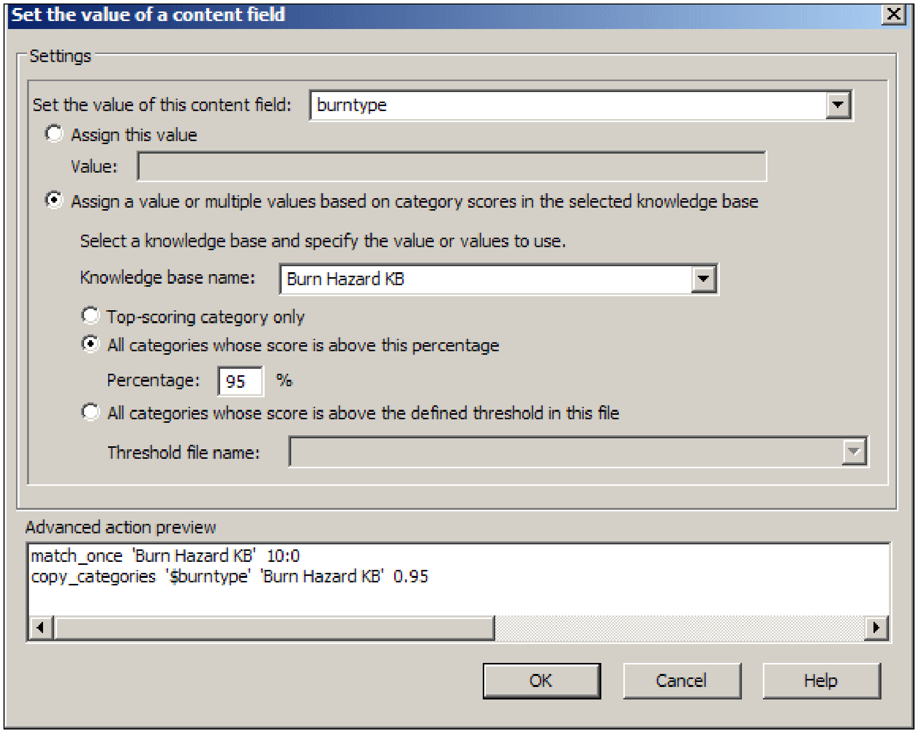
   d. In the Set the value of a content field window (Figure 9-11), complete these steps:

      i. Type the content field name. For this example, enter `burntype`.

      ii. Choose your knowledge base. For this example, select **Burn Hazard KB** as the knowledge base.

      iii. Choose **All categories whose score is above this percentage**. Type the confidence score. For this example, enter `95`.

   Figure 9-11 shows the new rule.



Set the value of a content field

Settings

Set the value of this content field: burntype

○ Assign this value

Value:

● Assign a value or multiple values based on category scores in the selected knowledge base

Select a knowledge base and specify the value or values to use.

Knowledge base name: Burn Hazard KB

○ Top-scoring category only

● All categories whose score is above this percentage

Percentage: 95 %

○ All categories whose score is above the defined threshold in this file

Threshold file name:

Advanced action preview

```
match_once  'Burn Hazard KB'  10:0
copy_categories  '$burntype'  'Burn Hazard KB'  0.95
```

Figure 9-11 Choosing burntype with a confidence score greater than 95%

8. Define a new rule to set the burntype field to the value `Unknown` when the confidence score is not above 95. Because you set the field in cases when the score is above 95, you know that, in all other cases, you have an empty burntype field. To define the rule, follow these steps:

   a. Create a rule. Type a new rule name. For this example, enter `Set BurnType Unknown`.

   b. Select **Trigger**. Click **Condition** and choose **Advanced**.

   c. Select **number = number**.

   d. Click the first number link. Choose **Size content field**. Click **Content field** and choose **burntype**.

   e. Click the second link and type `0`.

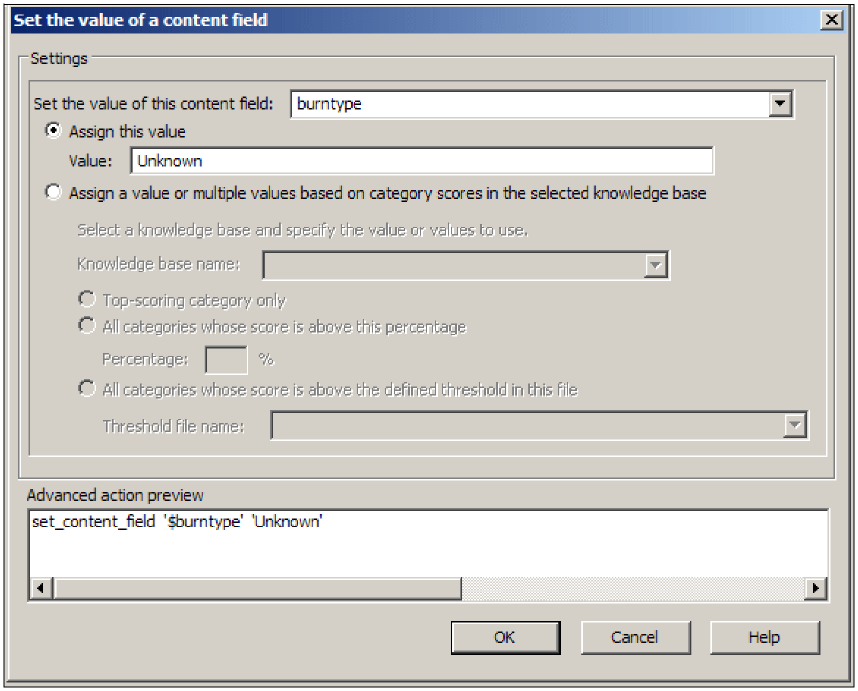
   f. Set the trigger size to `($burntype) = 0`.

   g. Select **Actions**. Click **Add** and choose **General actions**. Select **Set the value for a content field** and click **Next**.

   h. For the content field name, type `burntype`.

   i. Type the value `Unknown`.

   Figure 9-12 shows the new defined rule.



Set the value of a content field

Settings

Set the value of this content field: burntype

● Assign this value

Value: Unknown

○ Assign a value or multiple values based on category scores in the selected knowledge base

Select a knowledge base and specify the value or values to use.

Knowledge base name:

○ Top-scoring category only

○ All categories whose score is above this percentage

Percentage: %

○ All categories whose score is above the defined threshold in this file

Threshold file name:

Advanced action preview

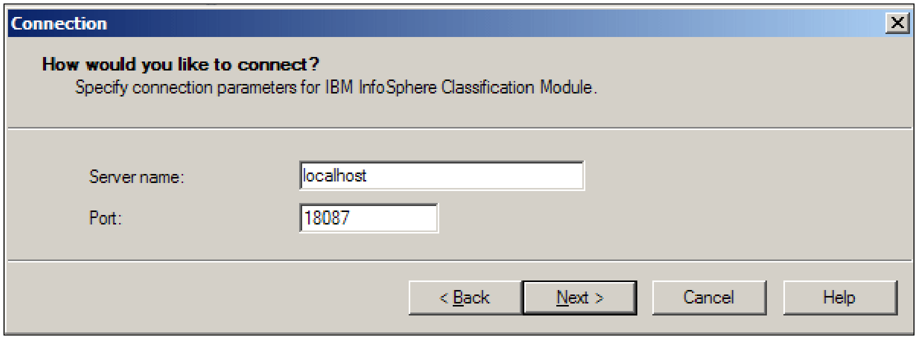
```
set_content_field '$burntype' 'Unknown'
```

Figure 9-12 Setting the burntype field to Unknown for low confidence scores

9.3.4 Deploying the knowledge base and decision plan

To use the knowledge base and decision plan you created, you have to deploy them to the Classification Module server:

1. Select **Project** → **Export**.
2. Select **Decision Plan**, and click **Next**.
3. Select **IBM Classification Module Server**, and click **Next**.
4. In the Connection window (Figure 9-13), specify the Classification Module Server machine name and port. Click **Next**.



Connection

How would you like to connect?

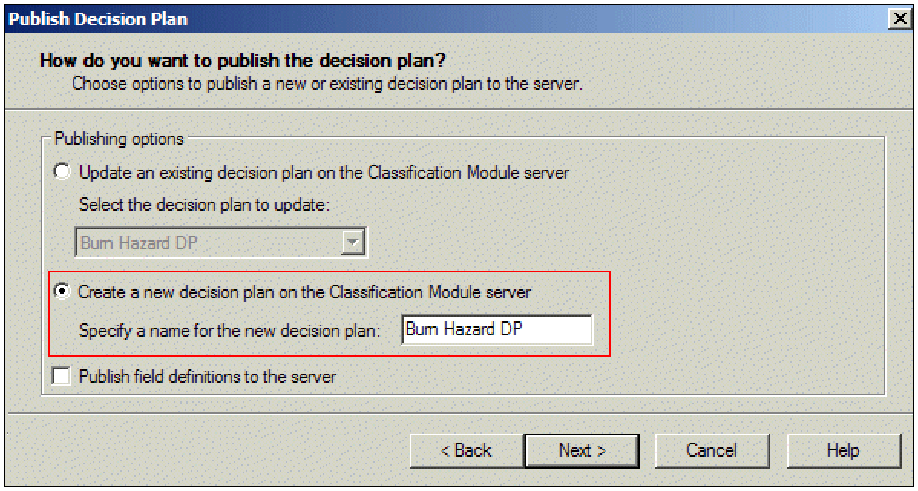
Specify connection parameters for IBM InfoSphere Classification Module.

Server name: localhost

Port: 18087

Figure 9-13 Workbench connecting to the Classification Module server

5. In the Publish Decision Plan window (Figure 9-14), select **Create a new Decision Plan on the Classification Module server**, and for Specify a name for the new decision plan, enter `Burn Hazard DP`. Then click **Next**.



Publish Decision Plan

How do you want to publish the decision plan?

Choose options to publish a new or existing decision plan to the server.

Publishing options

○ Update an existing decision plan on the Classification Module server

Select the decision plan to update: Bum Hazard DP

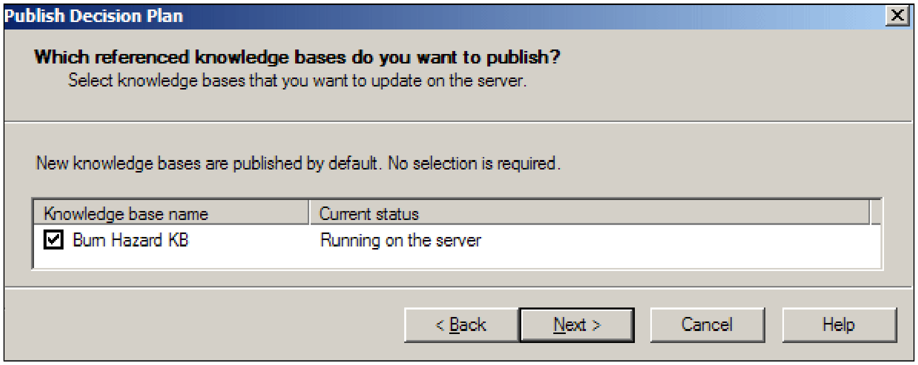
● Create a new decision plan on the Classification Module server

Specify a name for the new decision plan: Bum Hazard DP

□ Publish field definitions to the server

Figure 9-14 Publishing the decision plan to the Classification Module server

6. Select the knowledge base and click **Next** to publish the associated knowledge base (Figure 9-15).


Publish Decision Plan

Which referenced knowledge bases do you want to publish?
Select knowledge bases that you want to update on the server.

New knowledge bases are published by default. No selection is required.

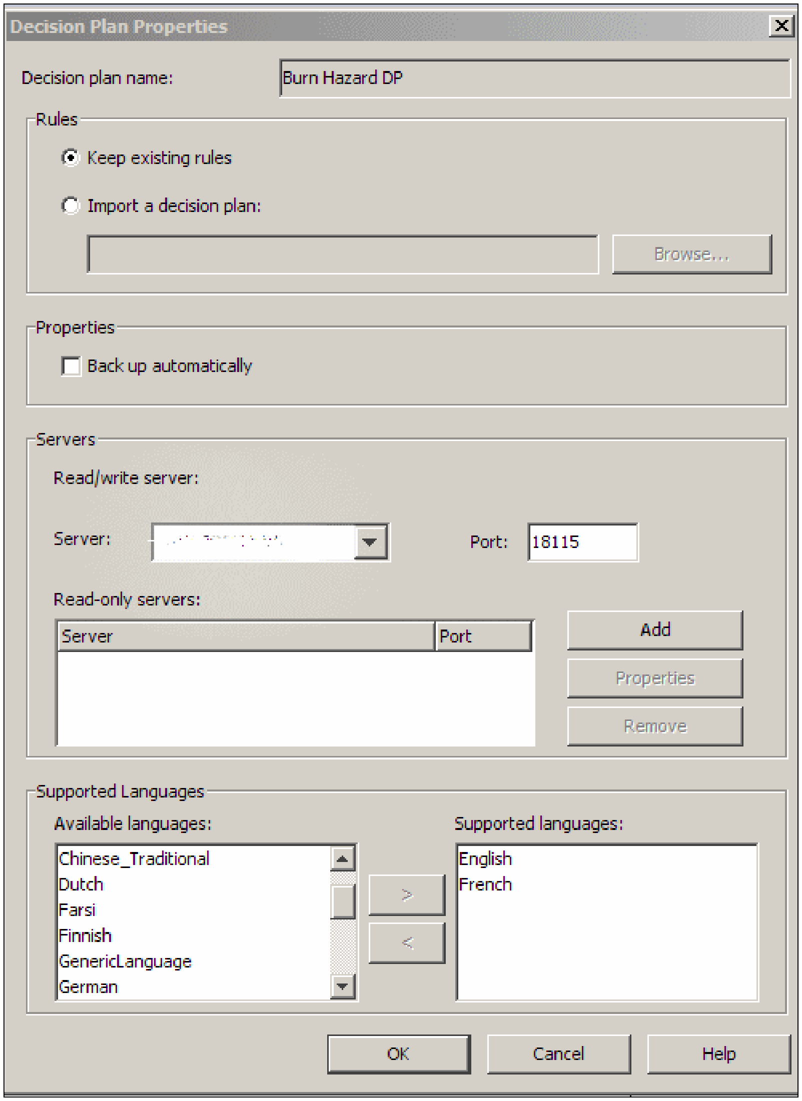
Knowledge base name	Current status
☑ Burn Hazard KB	Running on the server

Figure 9-15 Publishing the associated knowledge base to the Classification Module server

7. In the last window, click **Finish**.

Tip: Launch the Classification Module Management Console to check the status of the decision plan and knowledge base. Check that the list of languages for each knowledge base and decision plan corresponds to the languages that you foresee in your data (see Figure 9-16 on page 377).

Continuous operation: The decision plans and knowledge bases in Classification Module that you need to use with Content Analytics must run continuously. If you made changes to the system, such as adding a Classification Module catalog field, you must restart them accordingly.



Decision Plan Properties

Decision plan name: Burn Hazard DP

Rules
- ● Keep existing rules
- ○ Import a decision plan:

Properties
- □ Back up automatically

Servers

Read/write server:

Server:

Port: 18115

Read-only servers:

Server	Port

Supported Languages

Available languages:
Chinese_Traditional
Dutch
Farsi
Finnish
GenericLanguage
German

Supported languages:
English
French

Figure 9-16 Classification Module Management Console: Set languages

For more information about building a Classification Module knowledge base, see the technote *Building a knowledge base for IBM Classification Module V8.7* at the following address:

http://www-01.ibm.com/support/docview.wss?uid=swg27015916

Taxonomy Proposer

Classification Module provides clustering services with its *Taxonomy Proposer* application. You can use the data exported from Content Analytics and cluster it to discover groups of documents that share similar concepts and to define a new set of categories and a new taxonomy. You can then use the new categories to train a Classification Module knowledge base and use it to generate new, interesting facets in Content Analytics.

Content Analytics Clustering

The Classification Module clustering technology is integrated into the Content Analytics application. You can cluster a subset of the documents in the collection and deploy a categorization task to annotate the entire collection. See 8.3, “Document clustering” on page 343. You can also export a training set based on clustering suggestions (as facets).

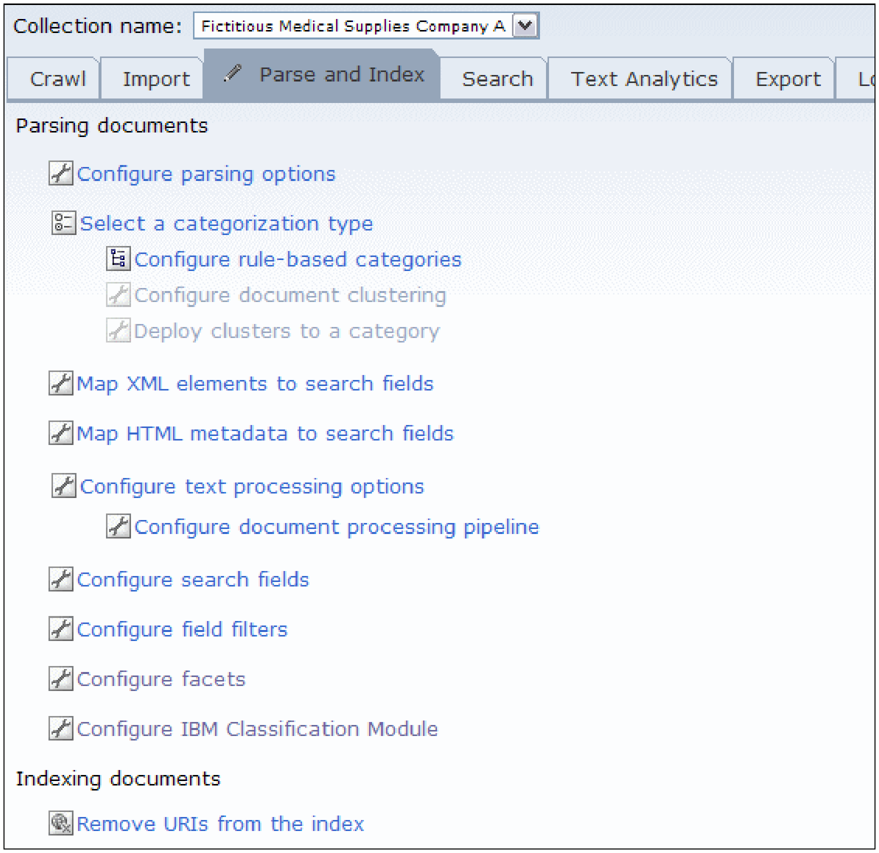
9.3.5 Configuring the Classification Module annotator

To configure the Classification Module, start by configuring the Classification Module decision plan:

1. Configure the Content Analytics parser to use Classification Module:

   a. From the administration console of Content Analytics, go to the Parse configuration of your collection, and click **Edit**. For this example, select the **Fictitious Medical Supplies Company A** collection.

b. On the **Parse and Index** tab, click **Configure Classification Module** (Figure 9-17).



Parsing documents

- Configure parsing options
- Select a categorization type
  - Configure rule-based categories
  - Configure document clustering
  - Deploy clusters to a category
- Map XML elements to search fields
- Map HTML metadata to search fields
- Configure text processing options
  - Configure document processing pipeline
- Configure search fields
- Configure field filters
- Configure facets
- Configure IBM Classification Module

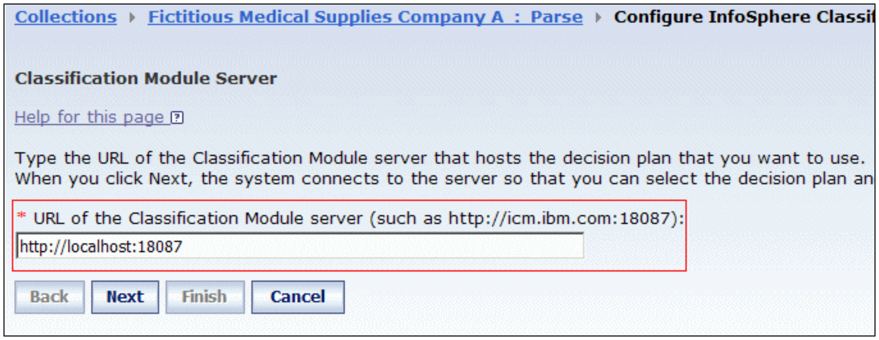
Indexing documents

- Remove URIs from the index

Figure 9-17 Selecting the configuration option on the Parse and Index tab

c. Start the Classification Module server.

d. In the Classification Module Server panel (Figure 9-18), type the URL of the Classification Module and click **Next**.



Collections ▸ Fictitious Medical Supplies Company A : Parse ▸ Configure InfoSphere Classi

Classification Module Server

Help for this page

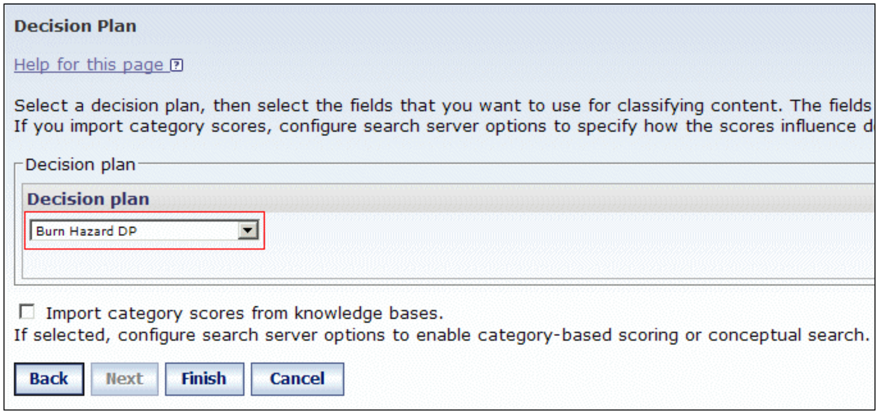
Type the URL of the Classification Module server that hosts the decision plan that you want to use. When you click Next, the system connects to the server so that you can select the decision plan an

* URL of the Classification Module server (such as http://icm.ibm.com:18087): http://localhost:18087

Figure 9-18 Classification Module URL setting

2. Associate a decision plan with your collection:

   a. In the Decision Plan panel (Figure 9-19), from the decision plan drop-down list, select the decision plan you want to associate with your collection. For this example, select **Burn Hazard DB**.



Decision Plan

Help for this page

Select a decision plan, then select the fields that you want to use for classifying content. The fields If you import category scores, configure search server options to specify how the scores influence d

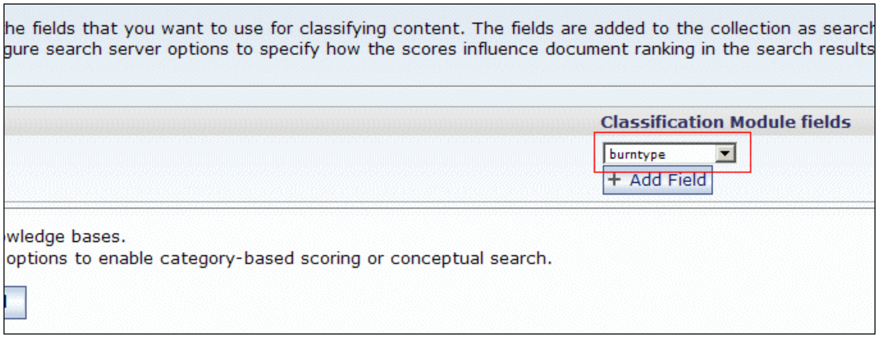
Decision plan

Decision plan: Burn Hazard DP

□ Import category scores from knowledge bases.
If selected, configure search server options to enable category-based scoring or conceptual search.

Figure 9-19 Selecting the decision plan

b. Map the Classification Module fields. In the Classification Module fields list (Figure 9-20), you see all the results fields that Classification Module can generate after classifying your data. In this case, select **burntype**, which is the only option available to select. In a case where you have more results fields, click **Add field**, and choose all the fields that you want to use in Content Analytics (for facets association and others).



he fields that you want to use for classifying content. The fields are added to the collection as search
gure search server options to specify how the scores influence document ranking in the search results

wledge bases.
options to enable category-based scoring or conceptual search.

Figure 9-20 Mapping the Classification Module field

c. Optional: Import category scores to enable category-based scoring or conceptual search.

3. Create a facet based on the fields generated by Classification Module:

   a. From the Content Analytics administration console, go to the parse configuration for your collection (for our example, Fictitious Medical Supplies Company A), and click **Edit**.

   b. Click **Configure facets** (see Figure 9-17 on page 379).

   c. Under the Hazards facet, add a facet:

      i. In the Add a facet panel, enter `Burns` as the new facet name and `BURNTYPE` for the facet path.

      ii. Click **Add** to complete the addition of the facet.

      iii. Select the **Add counts to parent facet** check box.

      iv. Clicking **Edit** next to the Fields mapping field to enable field mapping and choose **burntype**.

      v. Click **OK**.

> **Enabling field mapping:** The Classification Module fields are automatically enabled for field mapping.
>
> **After you make your changes**: You must deploy the resources for the changes that you made to be reflected in the index.

4. Enable the Classification Module annotator in the UIMA pipeline:
   a. From the administration console, go to the parse configuration for your collection and click **Edit**. For this example, go to Fictitious Medical Supplies Company A.
   b. Click **Configure document processing pipeline** (Figure 9-17 on page 379).
   c. Select **Classification Module Annotator** (Figure 9-21).

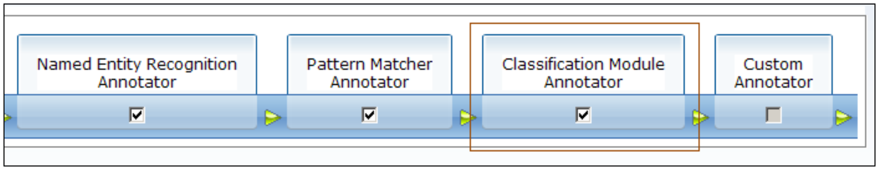

Named Entity Recognition Annotator
☑
Pattern Matcher Annotator
☑
Classification Module Annotator
☑
Custom Annotator
□

Figure 9-21 Selecting the Classification Module annotator

> **After you make changes:** You must either start or restart the parser for the changes to take effect. To apply the changes to documents in the index, reparse the documents and then rebuild the main index.

After you successfully reindex the content, you can review the facets that are generated by the Classification Module annotator in the text miner application. For this example, go to the Facets view and choose **Burn** to see the documents. You will find documents related to “chemical burns” that do not necessarily include the word “burn.” This result occurs because Classification Module looks at the entirety of the document to classify the results properly instead of just relying on a particular keyword.

> **Classification capability of Classification Module:** The Classification Module annotator uses the entire set of words in a document and can classify the documents related to a certain category based on understanding the full context.

9.4 Validation and maintenance of the Classification Module annotator

You use Classification Workbench to create and train a knowledge base and to create a decision plan. You also use Classification Workbench to tune and maintain your decision plan and knowledge base:

1. Import the data exported by Content Analytics. Then use Classification Module Workbench to train and fine-tune the knowledge base:

   – Use only part of the imported data to train the knowledge base and use the remaining part to analyze and tune it.

   – Use the Classification Module Workbench reports to assess the accuracy of the knowledge base.

   – Use the Classification Module Workbench reports to check the decision plan rules. In the Workbench decision plan projects, you have reports on the rules behavior for the overall content set. You can also trace for a specific content item that the rules have triggered and view their results by using the “Run item through decision plan” functionality. Go to the IBM InfoSphere Classification Module Information Center at the following address, and search on *decision plan analysis*:

     http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp

     > **More information:** For more tips about tuning Classification Module, see the IBM Redbooks publication *IBM Classification Module: Make It Work for You*, SG24-7707.

2. Import the XML with the suggested scores results to view the XML export data from Content Analytics that contains the analysis results.

   The Classification Workbench of Classification Module can generate accuracy reports from the data that contains the categories and the scores. For more information, search on the following topics in the IBM InfoSphere Classification Module Information Center:

   http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp

   – “Analyzing a knowledge base in production”
   – “Analyzing a decision plan in production”
   – “Sample XML output from saved analysis data”

9.4.1 Using the Classification Module sample programs

After deploying the decision plan and knowledge base to the Classification Module server, you can validate their behavior by invoking one of the sample applications that is installed with Classification Module. For example, you can use the `.\Samples\Java\JavaGUIDecide` application to connect to the decision plan that you created and published to the Classification Module server. You can introduce text or a document and observe the results.

> **Tip:** Consult the “Classification Module Tutorial” that is installed with Classification Module. To access the tutorial, select **Start** → **All Programs** → **IBM Classification Module 8.7** → **Classification Module Tutorial**.

9.4.2 Classification Module annotator validation techniques

The Classification Module annotator supports the following validation techniques:

- Verify the connection to the Classification Module server.
- Use the Classification Module samples to test that your decision plan acts as designed.
- Make sure that you reindex after deploying the Classification Module annotator and configuring the facets mapping.

9.5 Preferred practices

Text mining and gaining insight to your content is an iterative process. The following guidelines have been developed based on the field experiences of the authors of this book:

- Start your analysis with a small collection.
- Follow the normal procedure several times:
  - Crawl, parse and index, and inspect the content using the text miner application.
  - Define new dictionaries and inspect the content using the text miner application.
  - Define new pattern rules and inspect the content using the text miner application.
  - When you discover the need for a more sophisticated analysis, engage Classification Module.

- Training and tuning with Classification Module and defining decision rules can be a small iterative cycle in itself. Use the Classification Module Workbench to tune a knowledge base or refine the decision plan rules.
- Use the Taxonomy Proposer to refine your categories or to discover new categories.
- Use the Content Analytics clustering to discover classes of documents inside the collection and deploy a categorization task to annotate the entire collection accordingly.

Classification Module can invoke external hooks. When you need to engage Classification Module for more sophisticated text analysis, you can also use (if needed) its extensibility points to customize the text processing.

10

Importing CSV files, exporting data, and performing deep inspection

IBM Content Analytics supports importing comma-separated value (CSV) files into a collection so that you can quickly add content to the collection without setting up a crawler or accessing a content repository. It also provides a set of export capabilities to help you take advantage of the discoveries made by Content Analytics in your textual data with other tools.

This chapter provides information about CSV file import and the export features with several scenarios of how and why you might use export. It also explains deep inspection, which is a form of export that extends the text analysis capabilities of the text miner application to all of your data.

This chapter includes the following sections:

- Importing CSV files
- Overview of exporting documents and data
- Location and format of the exported data
- Common configuration of the export feature
- Monitoring export requests
- Enabling export and sample configurations

- Deep inspection
- Creating and deploying a custom plug-in

10.1 Importing CSV files

With the import functionality, you can add records in a CSV file to a text analytics collection so that the information is searchable when users work with the text miner application. The import functionality is easy to use. With it, you can quickly add content to the collection without setting up a crawler or accessing a content repository. You might decide to use this functionality if you are performing a proof of concept, wanting to perform content assessment on data that will not change over time, or are unable to directly connect to the content repository.

The CSV files need to conform to the RFC 4180 standard:

- The files must contain one or more records where each record is on a separate line and delimited by a line break (CR, LF, or CRLF).
- The fields within a record are delimited by a comma, white space, tab, or semicolon.
- The fields that contain escaped line breaks, quotation marks, or field delimiters must be enclosed in quotation mark characters.
- The file cannot contain more than 128 columns or be greater than 512 KB per record. Files greater than 128 MB are not supported for CSV file import when you select to upload a local file to the server.

In this scenario, you import a CSV file that contains two records. You import this file into the collection created in 4.3.2, “Creating a text analytics collection” on page 90.

1. Open the administration console and click the **Edit** icon for the Sample Text Analytics Collection.
2. Click the **Import** tab.

3. Click **Import CSV documents into the collection** (Figure 10-1).

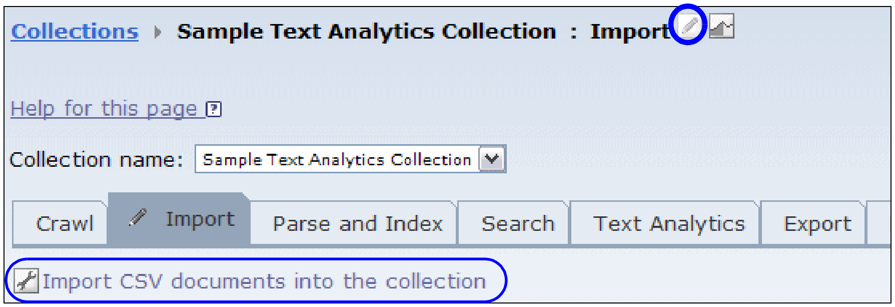

Collections
Sample Text Analytics Collection : Import
Help for this page
Collection name:
Sample Text Analytics Collection
Crawl
Import
Parse and Index
Search
Text Analytics
Export
Import CSV documents into the collection

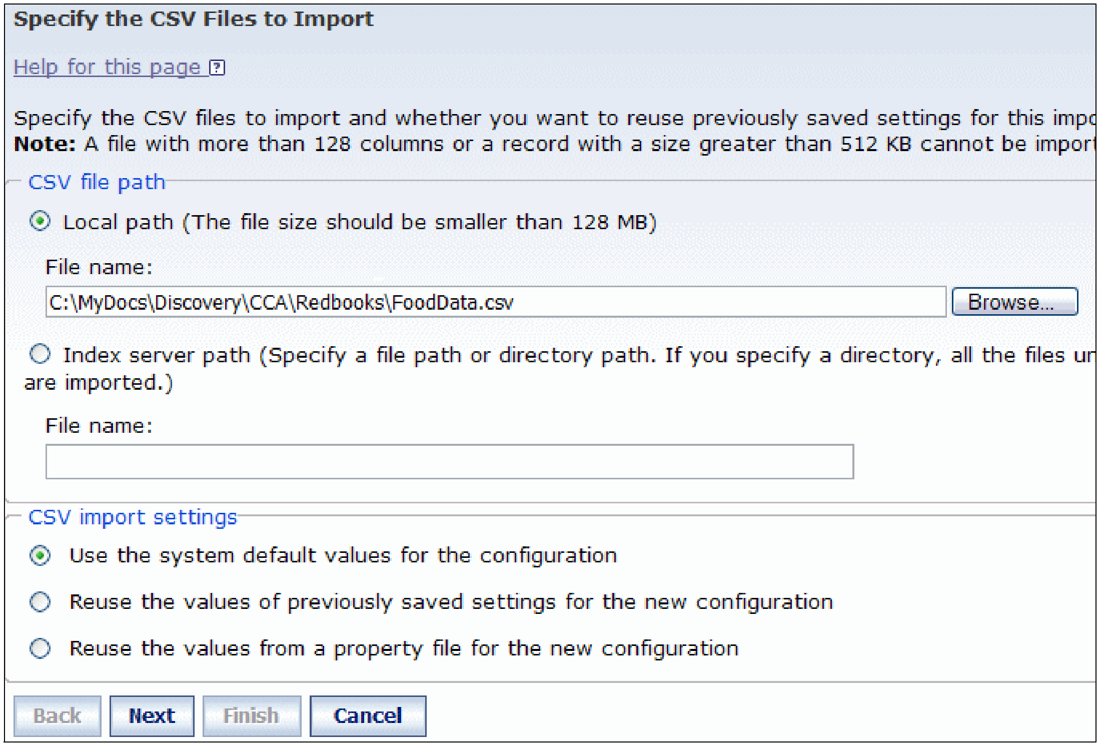
Figure 10-1 Selecting the 'Import CSV documents into the collection' option

4. Create a file that contains the following text and name it `FoodData.csv`:

```
Category,Product,Subcategory,Title,Body
Price,cookie,Price (general),Cookie - Price (general),I was charged
a higher price than what was advertised for my box of cookies.
Price,cookie,Price (general),Cookie - General Price,I was charged
twice for the same box of cookies.
```

5. In the Specify CSV Files to Import panel (Figure 10-2), for the File Name field under Local Path, click **Browse** and select the **FoodData.csv** document that you created in step 4.

> **Importing more than one CSV file:** To import more than one CSV file, select a directory path in the File Name field. This action results in adding all CSV files (`*.csv`) in the directory path to the index of the collection.



Specify the CSV Files to Import

Help for this page

Specify the CSV files to import and whether you want to reuse previously saved settings for this impo

Note: A file with more than 128 columns or a record with a size greater than 512 KB cannot be import

CSV file path

● Local path (The file size should be smaller than 128 MB)

File name: C:\MyDocs\Discovery\CCA\Redbooks\FoodData.csv

○ Index server path (Specify a file path or directory path. If you specify a directory, all the files ur are imported.)

File name:

CSV import settings

● Use the system default values for the configuration

○ Reuse the values of previously saved settings for the new configuration

○ Reuse the values from a property file for the new configuration

Figure 10-2 Specify the CSV File to Import panel

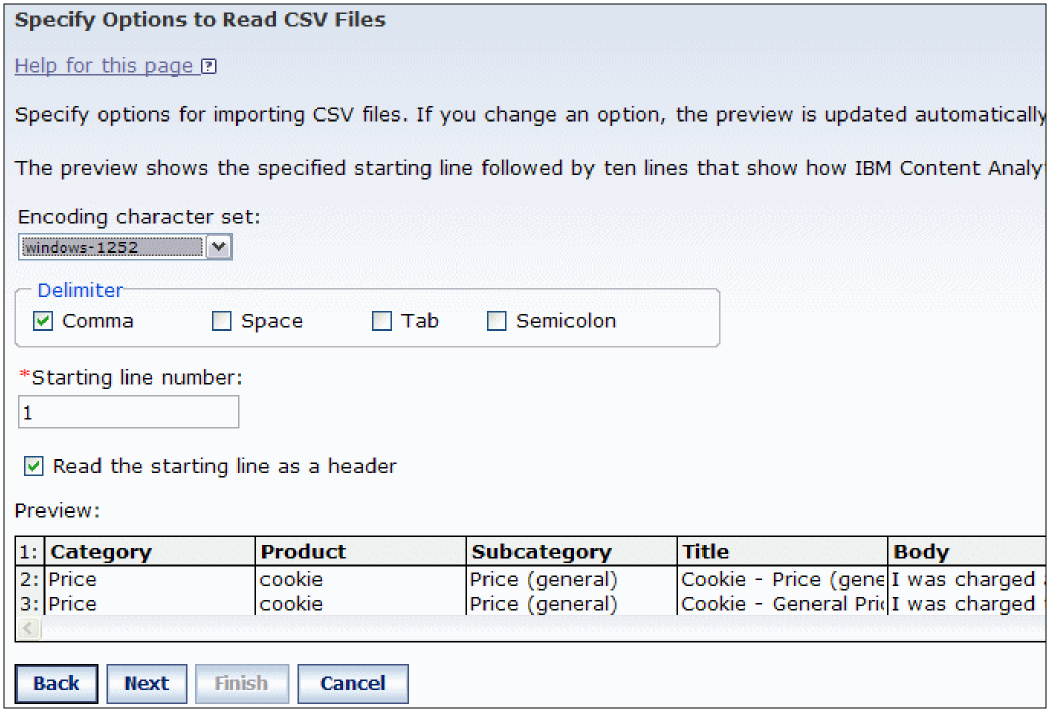
6. Select the **Use the system default values for the configuration** radio button and click **Next**.

Reimporting a CSV file: To reimport a CSV file and use the settings that you previously saved during import, select the **Reuse the values of previously saved settings for the new configuration** radio button. To reimport a CSV file and use the settings that you previously downloaded to a property file, select the **Reuse the values from a property file for the new configuration** radio button.

7. In the Specify Options to Read CSV Files panel (Figure 10-3), enter the field values defined in Table 10-1. Then click **Next**.

Table 10-1 Specify Options to Read CSV File field values

Field	Value
Encoding character set	windows-1202
Delimiter	**Comma**
Starting line number	1
Read the starting line as a header	(Select the check box)



Specify Options to Read CSV Files

Help for this page

Specify options for importing CSV files. If you change an option, the preview is updated automatically

The preview shows the specified starting line followed by ten lines that show how IBM Content Analy

Encoding character set:
windows-1252

Delimiter
☑ Comma □ Space □ Tab □ Semicolon

*Starting line number:
1

☑ Read the starting line as a header

Preview:

1:	Category	Product	Subcategory	Title	Body
2:	Price	cookie	Price (general)	Cookie - Price (gene	I was charged
3:	Price	cookie	Price (general)	Cookie - General Pri	I was charged

Figure 10-3 Specify Options to Read CSV Files panel

8. In the Specify the Columns to Import panel (Figure 10-4), complete these steps:

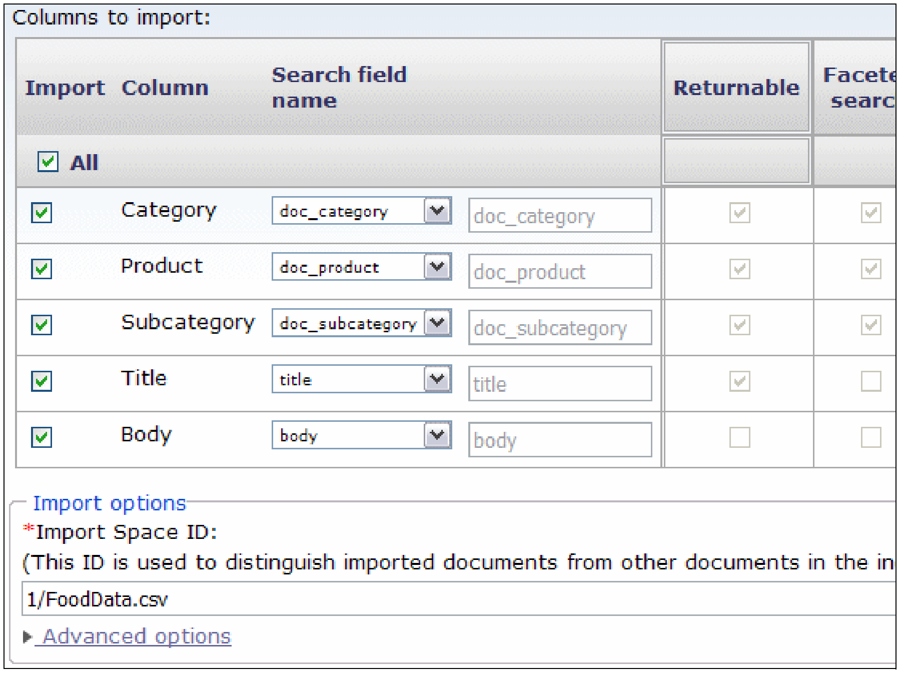
   a. Select the Search Field Name for the particular column based on the values listed in Table 10-2.

Table 10-2 Field values for the Specify the Columns to Import panel

Column	Search Field Name
Category	doc_category
Product	doc_product
Subcategory	doc_subcategory
Title	title
Body	body

   b. In the Import Space ID field, which must have a unique value, for this scenario, type `1/FoodData.csv`.

   c. Click **Next**.



Columns to import:

Import	Column	Search field name		Returnable	Facete searc
☑ All					
☑	Category	doc_category	doc_category	☑	☑
☑	Product	doc_product	doc_product	☑	☑
☑	Subcategory	doc_subcategory	doc_subcategory	☑	☑
☑	Title	title	title	☑	□
☑	Body	body	body	□	□

Import options

*Import Space ID:

(This ID is used to distinguish imported documents from other documents in the in

1/FoodData.csv

▸ Advanced options

Figure 10-4 Values used for the Specify the Columns to Import panel

Format of date and number values in data: If your data contains date values, set the Date Format and Time Zone fields (under the Advanced Options section) to the format of your date values. If your data contains number values, set the Number Format field (under the Advanced Options section) to the format of your number values. These formats are applied to all columns in the CSV file, which are mapped to date or number fields in the CSV file (that is, fields with “Parametric search” enabled).

9. In the Specify whether to save the current settings panel (Figure 10-5), complete these steps:

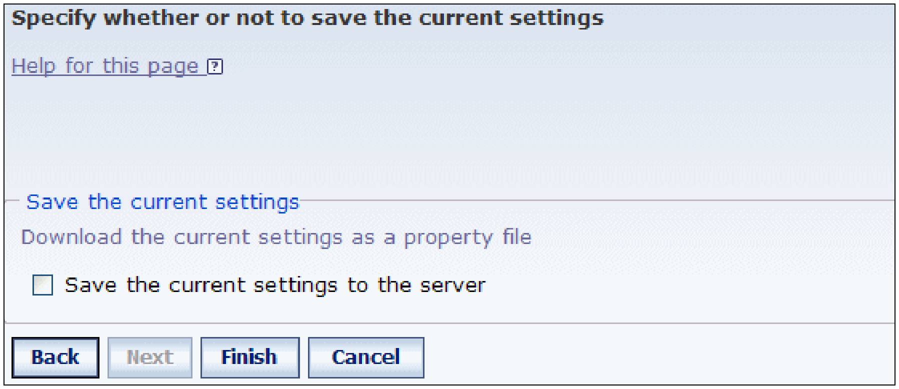
   a. Choose one of the following options for the current settings:

      - Save the configurations that you have defined as a property file by clicking **Download the current settings as a property file**.
      - Save the configurations to the server to be used at a later time by selecting **Save the current settings to the server**.

      In this scenario, do not save or download the import settings.

      Saving the imported settings: You can only save the import settings in the Specify whether to save the current settings panel. If you plan to import the same CSV file more than once or to reuse the import settings for other files, save the settings at this time.

   b. Click **Finish**.



Specify whether or not to save the current settings

Help for this page

Save the current settings

Download the current settings as a property file

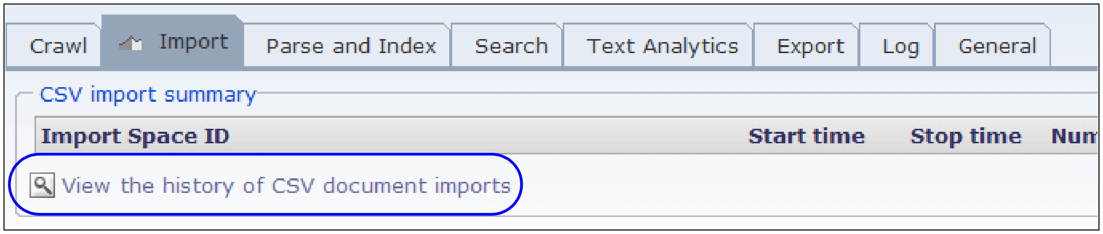
□ Save the current settings to the server

Figure 10-5 Specify whether to save the current settings panel

10.Now that you have set up the CSV file import, validate that the records were imported by reviewing the CSV document import history. To open the import history, click the **Monitor** icon.

11.Click the **Import** tab to change to the monitor mode.

12.Click **View the history of CSV document imports** (Figure 10-6).



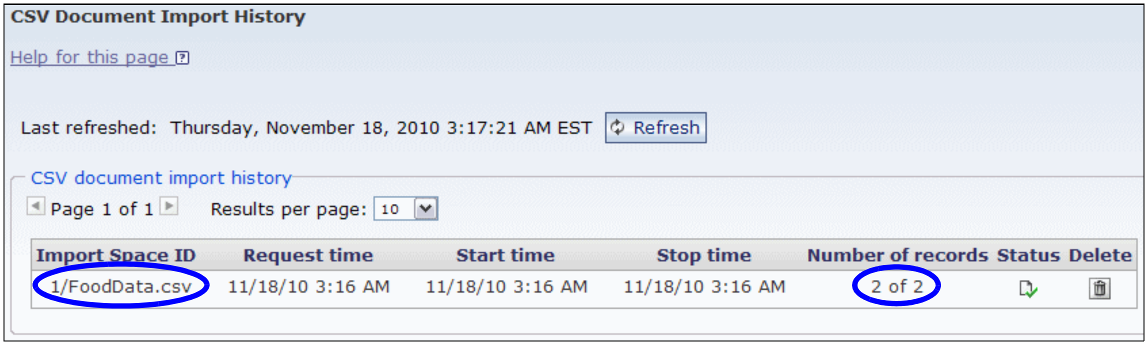
CSV import summary

Import Space ID	Start time	Stop time	Num

View the history of CSV document imports

Figure 10-6 CSV import summary panel

The row that contains the Import Space ID that matches the value used in step b on page 392 shows the results of the import. The value in the Number of records column indicates the number of records that were imported. In this scenario, two records were added to the index within the collection, as shown in Figure 10-7.



CSV Document Import History

Last refreshed: Thursday, November 18, 2010 3:17:21 AM EST

CSV document import history

Import Space ID	Request time	Start time	Stop time	Number of records	Status	Delete
1/FoodData.csv	11/18/10 3:16 AM	11/18/10 3:16 AM	11/18/10 3:16 AM	2 of 2		

Figure 10-7 History of the 1/FoodData.csv file import

Import CSV function: The import CSV file function does not keep track of modifications to the file. If you modify the CSV file, you must import the CSV file again which reimports all of the records in the CSV file into the collection.

Deleting the import history task: If you delete the import history task, the documents that were added to the index based on that particular import task are deleted from the index.

10.2 Overview of exporting documents and data

Content Analytics provides the powerful capability of analyzing structured and unstructured data (textual data) so that users can obtain actionable insight. In addition, with Content Analytics, users can also export data so that they can exploit the results of their analysis by using other applications such as data warehouse, business intelligence, or classification applications. Many IBM products can import and use this exported information, including IBM Content Collector, IBM Classification Module, and IBM Cognos Business Intelligence (BI).

Rationale for exporting data from Content Analytics

You might want to export data from Content Analytics for several reasons. For example, when two companies merge, both companies have data stored in different sources at different locations. In this situation, you can use Content Analytics to crawl data from various sources and then export the data to the file system. This exported data can be used later by Content Collector for archiving to a desired location.

Content Analytics can export data to a relational database in a star schema model. By using business intelligence or data warehouse applications, which access data from a relational database, analysts can gain a unique advantage of analyzing both structured and unstructured content together.

Content Analytics can connect to different data sources and collect large numbers of documents. Using the discovery capability of Content Analytics, customers can filter through this large volume of data and find relevant information by using Classification Module. Users can then export a subset of this data to be used by other software for usage such as generating reports.

Classification Module offers the ability to automatically and accurately classify documents into proper categories within Content Analytics. To use this capability, users must first export a sample set of data for each category to train the knowledge base in Classification Module. With Content Analytics, you can export the search result documents as sample data set to be used by Classification Module. For details about this use case, see Chapter 11, “Configuring annotators” on page 449.

Exporting points (stages) in Content Analytics

You can export data from Content Analytics during the following stages as illustrated in Figure 10-8:

- Export point 1: After documents are crawled.
- Export point 2: After documents are analyzed (processed and indexed).
- Export point 3: After a search is performed. You can then export the search result.

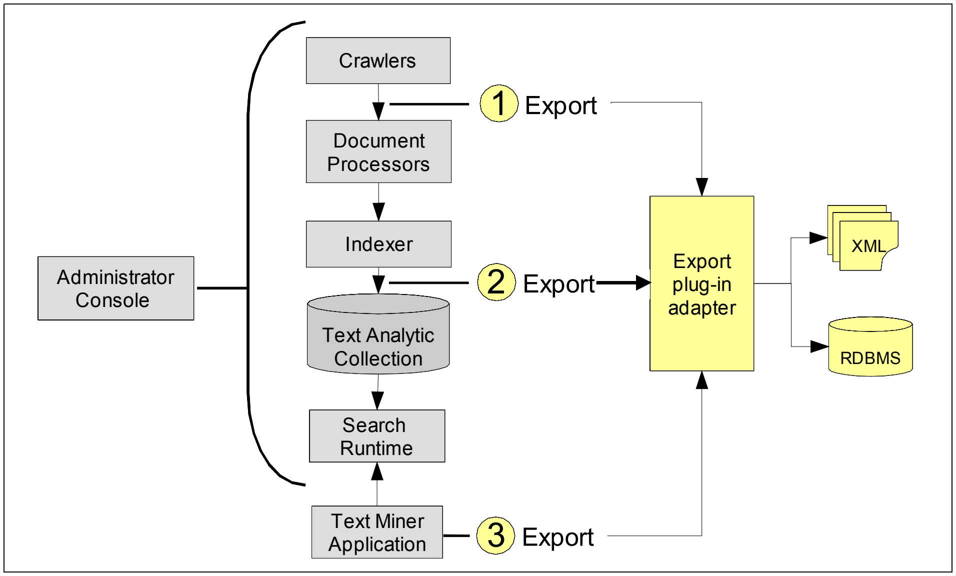


Figure 10-8 Export points in Content Analytics

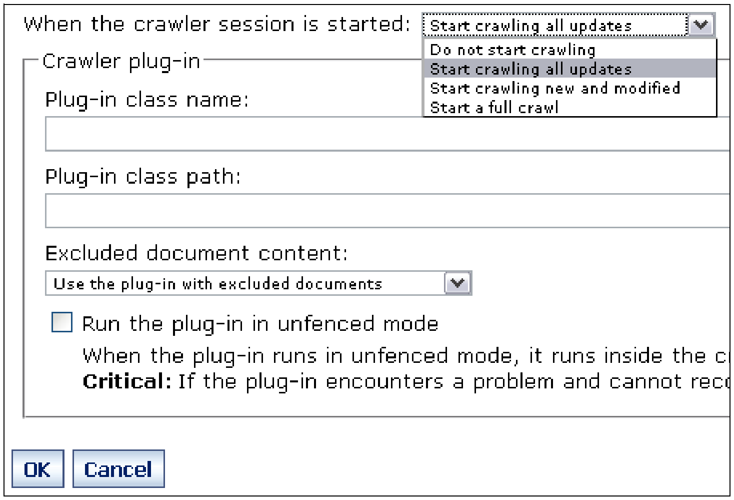
At each of these stages, you have the option to export data to either a file system or a relational database. Additionally, you can configure the deep inspection feature on a large text analytics collection to export the analysis results. For more information about the export options at each stage and deep inspection, see 10.6, “Enabling export and sample configurations” on page 408, and 10.7, “Deep inspection” on page 431.

Import feature: Content Analytics does not provide an import feature to add the previously exported data from Content Analytics into the same or different collection.

10.2.1 Crawled documents

Crawled documents are documents that have been retrieved from their data source but are not yet parsed or analyzed by Content Analytics. You can export them from Content Analytics. When you export crawled documents, you can choose to export metadata, binary content, or both. Additionally, you can export data so that it can be imported later by Content Collector.

The configuration of a crawler determines which crawled documents are to be exported. As shown in Figure 10-9, a crawler session can be configured to crawl data in different ways when it is started.



When the crawler session is started: Start crawling all updates

Crawler plug-in

Plug-in class name:

Plug-in class path:

Excluded document content: Use the plug-in with excluded documents

□ Run the plug-in in unfenced mode

When the plug-in runs in unfenced mode, it runs inside the c
Critical: If the plug-in encounters a problem and cannot rec

Figure 10-9 Configuration options for the crawler session

When you select **Start a full crawl** for the crawler, all the documents are recrawled, and thus all of the documents are exported.

When you select **Start crawling all updates** for the crawler, the crawler only retrieves new, modified, and deleted documents since the last crawl. In this case, Content Analytics exports new and modified data and information about the deleted data. For deleted documents, the value of the /Document@Type element in exported data is set to DELETED. When a document is deleted from the source, Content Analytics *does not* delete already exported data to synchronize with the source. In this case, you must be aware of the deleted documents and handle synchronization systematically or manually if required.

When you select **Start crawling new and modified data** for the crawler, the crawler only crawls new and modified data. The crawler does not check to see if any previously crawled documents were deleted. Therefore, Content Analytics exports only new and modified data. It does *not* export information about deleted documents because it does not have knowledge about deleted data.

10.2.2 Analyzed documents

Analyzed documents contain metadata, textual data, and any annotations that are added during the Unstructured Information Management Architecture (UIMA) document processing pipeline. You can export analyzed documents from Content Analytics. When you export analyzed documents, you have the option of exporting metadata and facets, Common Analysis Structure (CAS), extracted text, or both. You can use this option to perform advanced analysis on structured documents. For example, Content Analytics can export facets to a relational database in a star schema model so that it can later be used by IBM Cognos Business Intelligence to build meaningful reports.

Additionally, exported data at this stage contains a consistent and unified view of metadata from the different crawled data sources. Companies can often employ different systems to manage their information depending on the business need. For example, you might have active design or enhancement discussions stored in an IBM Lotus Notes database and have documents related to released products stored in a content management system such as IBM FileNet Content Manager.

Often fields that are semantically the same are named differently between data sources. In this scenario, the discussion documents in a Lotus Notes database might have a native field named “From” to identify the initiator of the discussion. In this same scenario, documents in the content management system might have the native field named “Author” to identify the creator of the document. Both of these fields represent the owner of the document. Assuming these native fields are mapped to a single search field named “Owner in Content Analytics,” they become normalized. When you export the analyzed documents, the values of the From and Author native fields are exported as values for the Owner field. This method relieves you of exported data from knowing the details of the individual data sources and allows normalized analysis of data from multiple disparate sources.

10.2.3 Search result documents

Search result documents are the set of documents returned for the query you execute by using the text miner application. You can export these documents from Content Analytics. You can also schedule an export of the search result documents on a recurring basis. By using the scheduling capability, you can periodically export the data of interest without having to manually mine the collection every time. For example, if new data is being added to the collection on a regular interval, you can use the *incremental export* feature where only documents added and updated after the last export are exported.

10.2.4 Exported data manifest

Depending on which stage you are at when you export your data, you can export different types of data, including metadata, binary content, CAS, and extracted text.

Metadata

Metadata entails the properties associated with documents. File size, date created, and author can be thought of as conventional metadata for files such as PDF files. Content Analytics also considers native fields either intrinsic to the document or managed externally from the document by the data source as metadata.

When you export analyzed documents, metadata includes facets populated by annotators in the UIMA processing pipeline. Additionally, if native fields are mapped to search fields, the exported metadata contains the names of search fields instead of the names of native fields. When exporting analyzed or search result documents, metadata includes only search fields that are configured as returnable.

Binary content

Binary content is the unstructured part of a document. Binary content is maintained in its native format such as a Microsoft Word document or PDF document.

Common Analysis Structure

When you export analyzed documents to a file system, you can also export the output of the UIMA document processing pipeline known as the CAS format. The CAS data is formatted as XML data and contains extracted text, annotations, facets, and other results of analysis. Exported CAS data to file system can also be used to validate the output of custom annotators. The CAS format conforms to the UIMA standard and is subject to change with future releases of UIMA.

Extracted text

Text extracted from the binary content is referred to as *extract text*. Extracted text is also referred to as *parsed content*. You can export some or all of these types of data at each stage of export.

Table 10-3 lists the configuration options at each stage. Even though you can configure the same options at different stages, the output at each stage can be different. For example, you can configure exporting metadata after both the crawled document and analyzed document stages. However, exported metadata for documents after the crawled document stage only contains conventional metadata. Exported metadata for analyzed documents also contains analyzed facets in addition to the conventional metadata.

Table 10-3 Export configuration options

Exported data	Crawled documents	Analyzed Documents	Search result documents
To file system or relational database	Yes	Yes	Yes
Metadata	Yes	Yes	Yes
Binary content	Yes		Yes
CAS		Yes	
Extracted text		Yes	Yes
For Content Collector integration	Yes	Yes	Yes
For Classification Module integration			Yes
Schedulable			Yes
Customize export	Yes	Yes	Yes

10.3 Location and format of the exported data

The type and format of data that is exported and the location to which the data is exported depends on the stage that you selected for export. This section explains where the documents are exported and the format in which they are exported.

10.3.1 Location of the exported data

When exporting data to a file system, you are required to provide a path to an existing directory for metadata and content. When the export service runs, a new folder is created under that path, except when exporting search result documents for Classification Module integration.

The name of the folder is based on the date and time that the export occurs. The name of the created folder is in the *yyyymmddhhmm* format. For example, if the export service starts on 30 March 2010 at 5:25 p.m., the folder name `201003301725` is created. After that, subdirectories are created for every 1000 documents exported that start with 0 and are incremented sequentially. The number 1000, indicating the number of documents in a folder, is a nonconfigurable parameter.

Example 10-1 shows the directory structure if you configure the export path of crawled documents as `C:\Export\CrawledDataExport`.

Example 10-1 Output for crawled documents into one directory

```
C:\Export\CrawledDataExport\201003301725\0
      0000.xml
      0000.dat
      0001.xml
      0001.dat
      .....
C:\Export\CrawledDataExport\201003301725\1
      0000.xml
      0000.dat
      0001.xml
      0001.dat
      .....
```

Additionally, when you configure the export options, you can provide different paths for the metadata and content. Example 10-2 shows a directory structure where metadata is exported into the `C:\Export\CrawledDataExport\metadata` directory, while content is exported into the `C:\Export\CrawledDataExport\content` directory.

Example 10-2 Output for crawled documents into separate directories

```
C:\Export\CrawledDataExport\metadata\201003301725\0
      0000.xml
      0001.xml
      .....
C:\Export\CrawledDataExport\metadata\201003301725\1
      0000.xml
      0001.xml
      .....

C:\Export\CrawledDataExport\content\201003301725\0
      0000.dat
      0001.dat
      .....
C:\Export\CrawledDataExport\content\201003301725\1
      0000.dat
      0001.dat
      .....
```

When you export search result documents for Classification Module integration only, the output files are created directly under the path provided in the configuration. Only two output files are created. One is the `catalog.xml` file. The other file is an XML file that contains metadata and the text combined. The file name consists of `results_<timestamp>.xml`. For example, if you enter the `C:\Export\SearchResultDataExport\ForICM\` path, the output has the following structure:

```
C:\Export\SearchResultDataExport\ForICM\
catalog.xml
results_1269593741.xml
```

10.3.2 Metadata format

Metadata is generally the structured part of a document and is often referred to using native fields. Metadata is exported in the XML format to a file system or mapped to relational database columns. The name of the native field is preserved in the XML file when exporting crawled documents. When Content Analytics analyzes documents, the native fields are mapped to the search fields, and the name of the search fields are used instead.

For example, a native field named *timestamp* can be mapped to the search field date in Content Analytics. When you export metadata for the analyzed and search result documents, the name of the search fields is used, not the name of the native fields. In this example, the exported metadata file contains the field date, not the time stamp. Furthermore, when exporting analyzed or search result documents, the metadata file only contains search fields that are configured to be returnable in the Content Analytics administration console.

The metadata format remains the same for Content Collector integration. However, the attributes and field names are converted to the XML element to allow XML metadata mapping in Content Collector.

Metadata file name

When a user configures Content Analytics to export metadata to a file system, the metadata is exported as XML files. The name of the first exported XML file begins with `0000.xml` and increments sequentially. For example, if the source file is `sample.doc`, the metadata file name is `0000.xml`. The XML file contains the original file name as a portion of the metadata.

Metadata in a relational database

Metadata is added to the database as columns of the table.

10.3.3 Binary content format

Binary content contains text or the unstructured part of the document and is exported in the original format such as Word or PDF.

When exporting crawled documents to a file system, by default, Content Analytics exports content as `.dat` files *and* preserves the binary format of the file. For example, you export a `.doc` source file in the `.dat` format and rename the `.dat` format to the `.doc` format. In this case, the resulting `.doc` file is the same as the original `.doc` file. However, when you export to a `.csv` file, the binary content of the document is not exported.

When you configure the export of crawled content for Content Collector integration, Content Analytics preserves the extension from the source (if available) and exports content with the original extension. For example, if the source document has the document `sample.doc` file, the exported content also has the `sample.doc` file name. For situations where Content Collector integration is enabled, but the source document does not provide a file extension, the document is exported as a `.dat` file.

Binary content file name

When you configure Content Analytics to export content to a file system, the name of the XML file begins with `0000.dat` and increments sequentially. For example, if the source file is the `sample.doc` file, the metadata file name is `0000.dat` or `0000.doc` if Content Collector integration is enabled.

Binary content in relational database

Binary content is stored in the relational database as a binary large object (BLOB).

10.3.4 Common Analysis Structure format

CAS is a data structure for representing information that is gathered during the analysis of document such as annotations, tokens, and facets. When you export the CAS format to a file system, the data is exported in the XMI format as `.xmi` files. The CAS format conforms to UIMA standards and is subject to change with future releases of UIMA.

When Content Collector integration is enabled, the format of the CAS file does not change.

Common Analysis Structure file name

The name of the XMI file begins with `0000.xmi` and increments sequentially until the name reaches `9999.xmi`. When more than 10,000 documents are exported, a new folder is created with file names beginning with `0000.xmi`.

Common Analysis Structure in relational database

CAS export to relational database is not supported.

10.3.5 Extracted text format

Extracted text is the unstructured part of the document. The text contains extracted characters from binary content.

When you export extracted text to a file system, the data is exported in the same XML file that contains the metadata. It is included in the `<content></content>` element. When Content Collector integration is enabled, the format of the extracted text does not change.

Extracted text file name

Extracted text is part of the metadata. See “Metadata file name” on page 403 for more information.

Extracted text in a relational database

When you export analyzed documents to a relational database, the extracted text is exported as a character large object (CLOB) into a column.

10.4 Common configuration of the export feature

The three export stages share common configuration features when exporting crawled, analyzed, or searched documents. This section provides details about these common configuration options.

If you run into problems after configuring the export feature, see 16.7, "Export-related troubleshooting" on page 624, for troubleshooting guidance on some of the common problems.

10.4.1 Document URI pattern

Content Analytics supports limiting which documents are exported based on the composition of the Uniform Resource Identifier (URI) of the document. You can enter a list of regular expression patterns as a value of this configuration property, and Content Analytics only exports those documents whose URI matches one of these patterns. For example, if you provide the following example as a value for the Document URI patterns field, only PDF and Word documents are exported:

```
.*.pdf
.*.doc
```

10.4.2 Exporting XML attributes and preserving file extensions

If you plan to import into Content Collector the crawled, analyzed, or search result documents that are exported from Content Analytics, you must select the **Use field name or facet path as XML element** check box. By selecting this option, you can export the XML attributes and field names as elements so that the elements can be used during metadata mapping in the Content Collector configuration.

This option also preserves the extension of native file names when the binary content is exported. Although this configuration is required to enable Content Collector integration, you can use it for other reasons. For example, you can enable this option to preserve the extension of the binary content in the exported data.

10.4.3 Adding exported documents to the index

When you configure export options for crawled or analyzed documents, you can select the **Do not add the exported documents to the index** check box. If the purpose of using Content Analytics is to collect data or collect parsed information to be redirected to a different destination, you can save hardware resources by not building an index of the data in Content Analytics. When you select this option, the Text Mining application for the collection is not available.

10.4.4 Exporting information about deleted documents

By default, information about deleted documents is also exported when you export crawled or analyzed documents. To disable this option, when you configure export options for crawled or analyzed documents, select the **Do not export information about deleted documents** check box.

10.4.5 Scheduling

For search-result documents, export can be scheduled to start at a later time. When scheduling an export request, you can specify when the export operation is to start and how often it must run. For example, you can schedule an export request to run at off-peak hours without impacting the production time search capability. Additionally, you can disable specific export request or even delete it if you no longer need to export the data.

Incremental export

Incremental export means exporting only new documents that are added after the last export. When you have a dynamic collection where data is being added on regular basis, you can export documents on an incremental basis to keep the data accurate and up-to-date.

Custom schedule

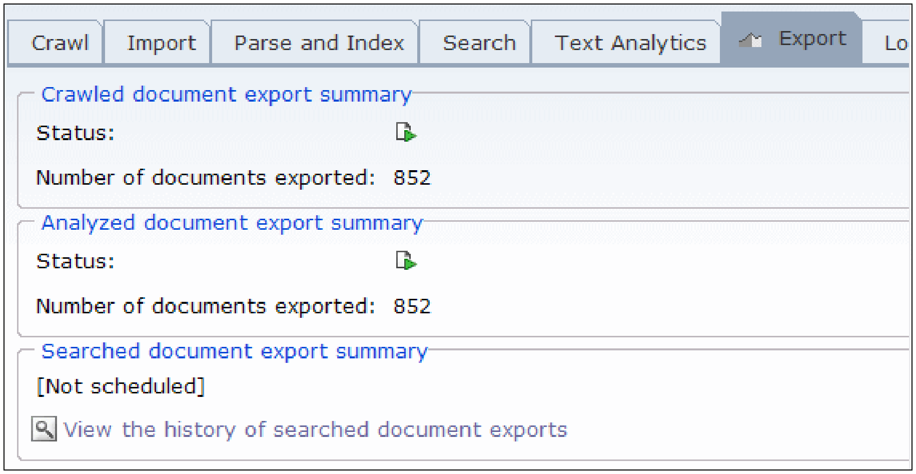
You can configure the export request to run on a general schedule specified for all the requests, or you can customize a schedule for each discrete request to run at a specific time.

Configuring and enabling scheduling: After the export request is submitted through the text miner application, the administrator *must* configure and enable the schedule for the export request for Content Analytics to export documents at the scheduled time:

1. Enable the export feature for searched documents by using the administration console.
2. Using the text miner application, perform a search and export the search result. For the export option, select the schedulable option.
3. In the administration console, configure and enable the schedule for the export request in step 2.

10.5 Monitoring export requests

Content Analytics provides a monitoring capability to help you see the status of export requests for crawled, analyzed, and searched data. Figure 10-10 shows an example export monitor page that provides a summary of all the crawled and analyzed documents.



Crawled document export summary

Status:

Number of documents exported: 852

Analyzed document export summary

Status:

Number of documents exported: 852

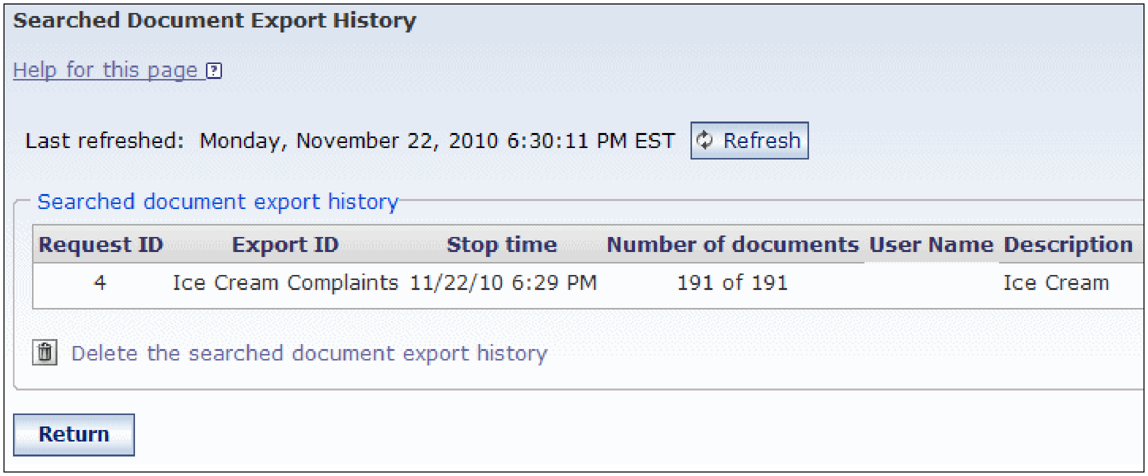
Searched document export summary

[Not scheduled]

View the history of searched document exports

Figure 10-10 Export Monitor view

The Export Monitor view has a link to the summary of export requests for the searched document. Figure 10-11 shows an example of the search result summary window.



Searched Document Export History

Last refreshed: Monday, November 22, 2010 6:30:11 PM EST

Searched document export history

Request ID	Export ID	Stop time	Number of documents	User Name	Description
4	Ice Cream Complaints	11/22/10 6:29 PM	191 of 191		Ice Cream

Delete the searched document export history

Figure 10-11 Searched Document Export History window

To view the export monitoring tool, follow these steps:

1. From the administration console, click **Collections** in the toolbar.
2. In the Collections view, locate the collection that you want to edit, and click the **Monitor** icon (Figure 10-12).

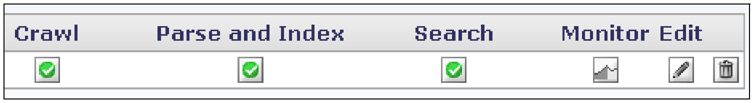


Figure 10-12 Collections view with the editing and monitoring options

3. Click the **Export** tab, and you see a page similar to the one in Figure 10-10 on page 407.
4. Optional: Click the **View the history of searched document exports** link to see all the export requests that are made for the search result document.

10.6 Enabling export and sample configurations

Before you begin the configuration for exporting data, consider these questions:

- What data must be exported: crawled, analyzed, or search result?
- Where will you export the data: the file system or a relational database?
- What metadata or fields do you want to include in the exported result?
- Which application is using the exported data: Content Collector, Classification Module, or another product?
- Is this a one-time export or a recurring export task?

As a first step, you can create a collection with a sample set of data, configure the desired export option, and then validate that the output to ensure that it is what you expect. For example, you can ensure that the metadata fields you want are present in the output.

This section takes you through the following export scenarios:

- Exporting crawled documents to a file system for Content Collector

  This scenario shows how to configure Content Analytics to export data to the file system for Content Collector usage.

- Exporting analyzed documents to a relational database

  This scenario shows how to create a proper database mapping file and configuring Content Analytics to export data to a relational database.

- Exporting search result documents to the file system for Classification Module

  This scenario shows how to configure the export of search result documents to be used by Classification Module.

- Exporting search results to CSV files

  This scenario shows how to configure the export of the search results as CSV files.

These scenarios are configured by using the collection created in Chapter 4, “Installing and configuring IBM Content Analytics” on page 71.

The Export icon

To export search results from the text miner application, click the **Export** icon ().

10.6.1 Exporting crawled documents to a file system for Content Collector

Content Analytics supports the crawling of over 25 types of enterprise data sources. Many applications can use this robust crawling feature of Content Analytics by exporting crawled results for usage. Content Collector is one such IBM product that can use Content Analytics to crawl the enterprise and archive the crawled content.

This section explains how to configure Content Analytics to export metadata and content to the file system. Content Collector uses the format of the exported data. The first step is to configure and run the export, and the second step is to validate that the data is correctly exported.

Configuring and running an export

To configure an export, follow these steps:

1. From the administration console, click **Collections** in the toolbar.
2. In the Collections view, locate the collection that you want to edit, and click the **Edit** icon (Figure 10-13).

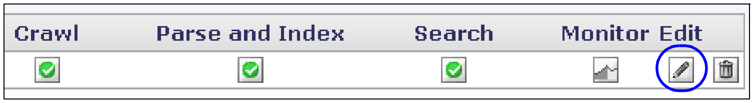
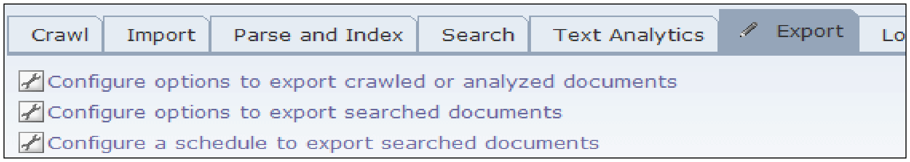


Figure 10-13 Collections view showing the editing and monitoring options

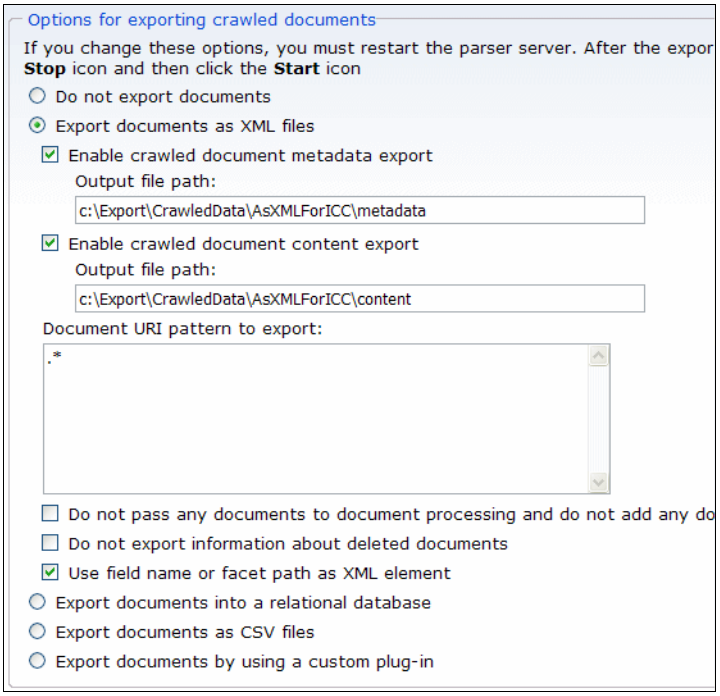
3. Click the **Export** tab (Figure 10-14), and click the **Configure options to export crawled or analyzed documents** link.



Configure options to export crawled or analyzed documents
Configure options to export searched documents
Configure a schedule to export searched documents

Figure 10-14 Export tab in edit mode

4. In the Options for exporting crawled documents section of the Crawled document export options window (Figure 10-15 on page 411), complete these steps:
   a. Select **Export documents as XML files**.
   b. Select the **Enable crawled document metadata export** and **Enable crawled document content export** check boxes.
   c. For the Output file path fields, enter the paths of the *existing* directory. You can provide the same path for both content and metadata or separate paths for each field.
   d. Select the **Use field name or facet path as XML element** check box, which is *crucial* for exported documents to be consumable by Content Collector.
   e. Click **OK**.



Options for exporting crawled documents

If you change these options, you must restart the parser server. After the expor
Stop icon and then click the **Start** icon

○ Do not export documents

● Export documents as XML files

☑ Enable crawled document metadata export

Output file path:

c:\Export\CrawledData\AsXMLForICC\metadata

☑ Enable crawled document content export

Output file path:

c:\Export\CrawledData\AsXMLForICC\content

Document URI pattern to export:

```
.*
```

☐ Do not pass any documents to document processing and do not add any do

☐ Do not export information about deleted documents

☑ Use field name or facet path as XML element

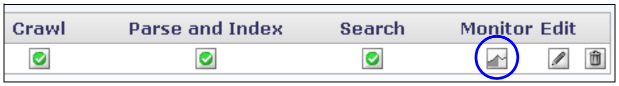
○ Export documents into a relational database

○ Export documents as CSV files

○ Export documents by using a custom plug-in

Figure 10-15 Crawled document export configuration

5. Click **Collections** in the toolbar.
6. In the Collections view (Figure 10-16), click **Monitor** for the collection.

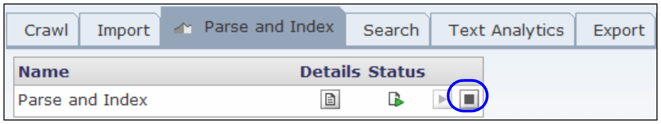


Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓		

Figure 10-16 Collections view with editing and monitoring options

7. Restart the document processor service. To restart the service:

   a. Click the **Parse and Index** tab (Figure 10-17).
   b. Click **Stop**.

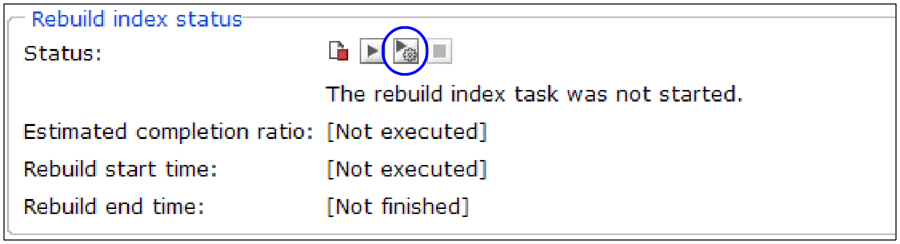


Name	Details	Status
Parse and Index		

Figure 10-17 Parse and index tab

   c. When the service is stopped, click **Start**.

If you have already built a collection, you must rebuild it after making this configuration change. To rebuild the index, click **Details**, and then click **Restart a full index build** (circled in Figure 10-18).



Rebuild index status

Status:

The rebuild index task was not started.

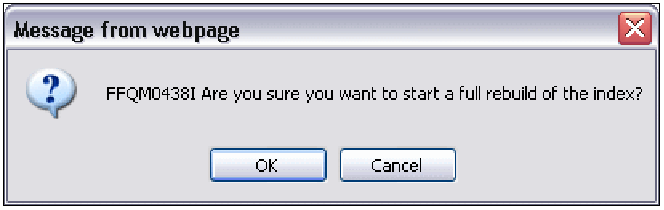
Estimated completion ratio: [Not executed]

Rebuild start time: [Not executed]

Rebuild end time: [Not finished]

Figure 10-18 Building a full index

8. In the confirmation message window that opens (Figure 10-19), click **OK**.

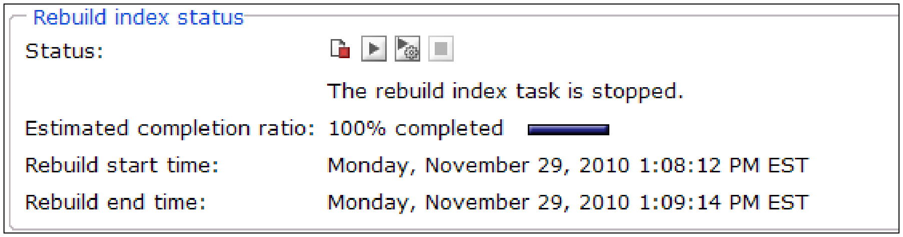


Message from webpage

FFQM0438I Are you sure you want to start a full rebuild of the index?

Figure 10-19 Full rebuild confirmation message window

9. Wait for the rebuild index to complete (Figure 10-20).



Rebuild index status

Status:

The rebuild index task is stopped.

Estimated completion ratio: 100% completed

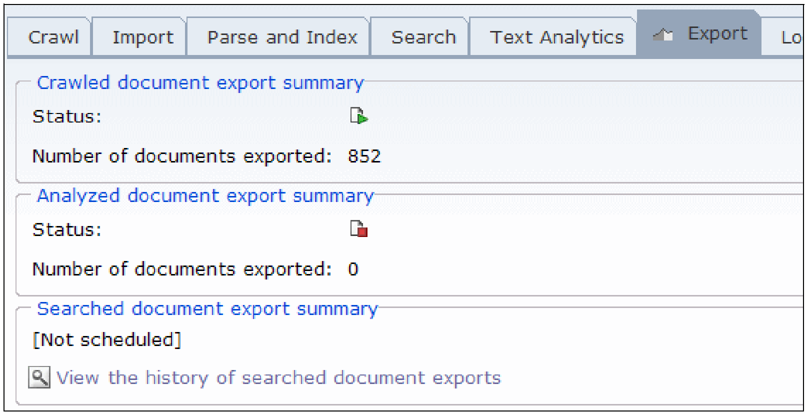
Rebuild start time: Monday, November 29, 2010 1:08:12 PM EST

Rebuild end time: Monday, November 29, 2010 1:09:14 PM EST

Figure 10-20 Completed rebuilding index process

Validating the result crawled documents export

Using export monitoring in the administration console, validate that the number of documents exported (Figure 10-21) is the same as the number of documents crawled.



Crawled document export summary

Status:

Number of documents exported: 852

Analyzed document export summary

Status:

Number of documents exported: 0

Searched document export summary

[Not scheduled]

View the history of searched document exports

Figure 10-21   Crawled document export summary

Verify that the directory that you specified as the output path to verify that data is being exported. Example 10-3 shows the format of the output file for metadata.

Example 10-3   Metadata file exported to a file system for Content Collector integration

```
<?xml version="1.0" encoding="UTF-8"?>
<Document Id="file:///C:/IBM/es/samples/firststep/data/xml/xml-data/00000851.xml"
Type="NORMAL">
  <Content>
    <Path>c:\Export\CrawledData\AsXMLForICC\content\20100401025416\0\0000.xml</Path>
    <Directory>c:\Export\CrawledData\AsXMLForICC\content\20100401025416\0</Directory>
    <Name>0000.xml</Name>
    <Truncated>false</Truncated>
  </Content>
  <Metadata>
    <Fields>

<Directory><![CDATA[C:\IBM\es\samples\firststep\data\xml\xml-data]]></Directory>
      <FileName>00000851.xml</FileName>
      <Extension>.xml</Extension>
      <ModifiedDate>1250055768000</ModifiedDate>
      <FileSize>453</FileSize>
      <Title>00000851.xml</Title>
    </Fields>
    <Facets></Facets>
  </Metadata>
</Document>
```

Example 10-4 shows a sample of the metadata file, which is also exported to the file system, *without* Content Collector integration enabled. Many of the attributes in Example 10-4 are transformed into XML elements, as shown in Example 10-3 on page 413. For example, look for the following line in Example 10-4:

```
<Field Name="__$FileName$__">00000851.xml</Field>
```

This line is transformed to the following lines in Example 10-3 on page 413 when Content Collector integration is enabled:

```
<Fields>
   <FileName>00000851.xml</FileName>
```

This switch allows seamless metadata mapping in Content Collector.

Example 10-4 Metadata file exported to a file system when Content Collector is disabled

```
<?xml version="1.0" encoding="UTF-8"?>
<Document Id="file:///C:/IBM/es/samples/firststep/data/xml/xml-data/00000851.xml"
Type="NORMAL">
  <Content Truncated="false"
Path="c:&#x5C;Export&#x5C;CrawledData&#x5C;AsXML&#x5C;content&#x5C;20100401030225&#x5
C;0&#x5C;0000.dat" Encoded="false"></Content>
  <Metadata>
    <Fields>
      <Field
Name="__$Directory$__">C:&#x5C;IBM&#x5C;es&#x5C;samples&#x5C;firststep&#x5C;data&#x5C
;xml&#x5C;xml-data</Field>
      <Field Name="__$FileName$__">00000851.xml</Field>
      <Field Name="__$Extension$__">.xml</Field>
      <Field Name="__$ModifiedDate$__">1250055768000</Field>
      <Field Name="__$FileSize$__">453</Field>
      <Field Name="__$Title$__">00000851.xml</Field>
    </Fields>
    <Facets></Facets>
  </Metadata>
</Document>
```

10.6.2 Exporting analyzed documents to a relational database

Before you begin exporting data to a relational database, you can export a small subset of data as XML to a file system for the following reasons:

- To see the type of data that is being exported as fields, facets, and metadata. For any native fields that you map to the search fields, the names of the search fields are displayed in the exported data.
- To determine the data that needs to be inserted into the database. For example, when you export analyzed documents, no binary data is exported. In this case, you must modify the default configuration. Similarly, you can modify the configuration to remove any fields or facets that you do not want to needlessly insert into the database.

Content Analytics uses the field configuration that you set for the export to automatically insert exported data into a relational database. Content Analytics inserts the data into star-schema tables.

Exporting analyzed documents to a relational database requires the following tasks, which are explained in the following sections:

1. Configuring the database export information
2. Running the export
3. Validating the results of the analyzed documents export

Configuring the database export information

The first step in exporting data to a relational database is to configure the database export information. You must define how the metadata of a document is mapped to the columns of tables in the database. The database information includes connection and configuration information about the database.

To set up the database export configuration information for exporting analyzed documents, follow these steps:

1. From the administration console, click **Collections** in the toolbar.
2. In the Collections view, locate the collection you want to edit and click **Edit**.
3. Select the **Export** tab (Figure 10-22), and click the **Configure options to export crawled or analyzed documents** link.

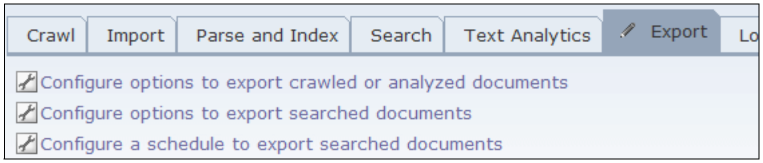
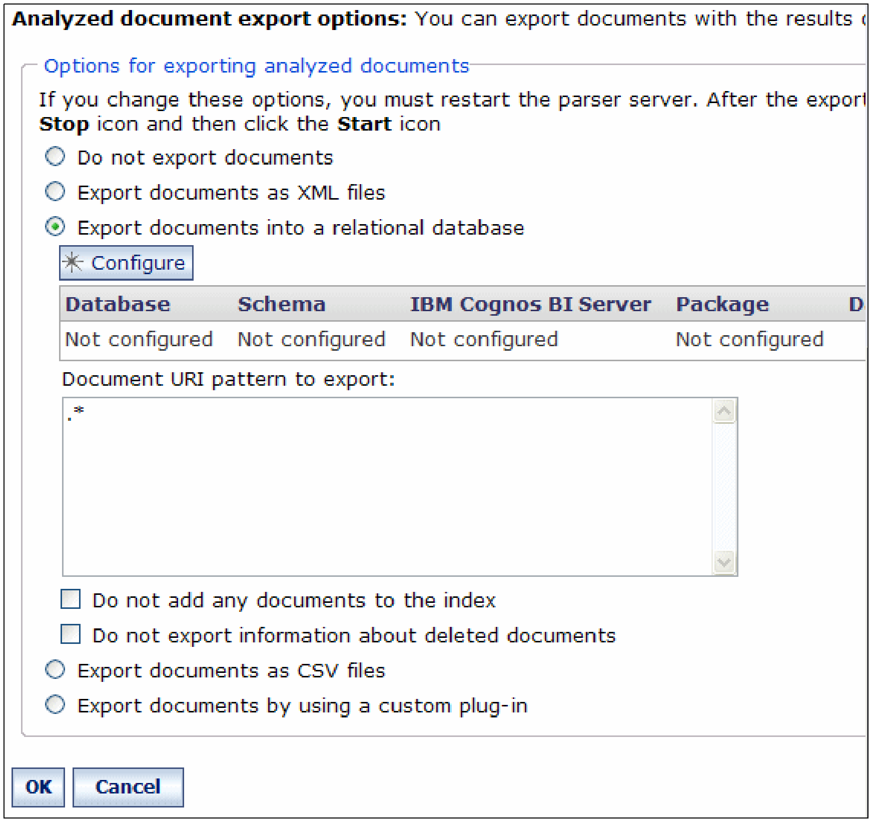

Crawl
Import
Parse and Index
Search
Text Analytics
Export
Lo
Configure options to export crawled or analyzed documents
Configure options to export searched documents
Configure a schedule to export searched documents

Figure 10-22 Export tab in edit mode

4. In the Export Searched Documents panel (Figure 10-23), follow these steps:

   a. Under Analyzed document export options, select the **Export documents into a relational database** option.

   b. Click **Configure**.



Analyzed document export options: You can export documents with the results

Options for exporting analyzed documents

If you change these options, you must restart the parser server. After the export
Stop icon and then click the **Start** icon

○ Do not export documents

○ Export documents as XML files

● Export documents into a relational database

Configure

Database	Schema	IBM Cognos BI Server	Package	D
Not configured	Not configured	Not configured	Not configured	

Document URI pattern to export:

```
.*
```

□ Do not add any documents to the index

□ Do not export information about deleted documents

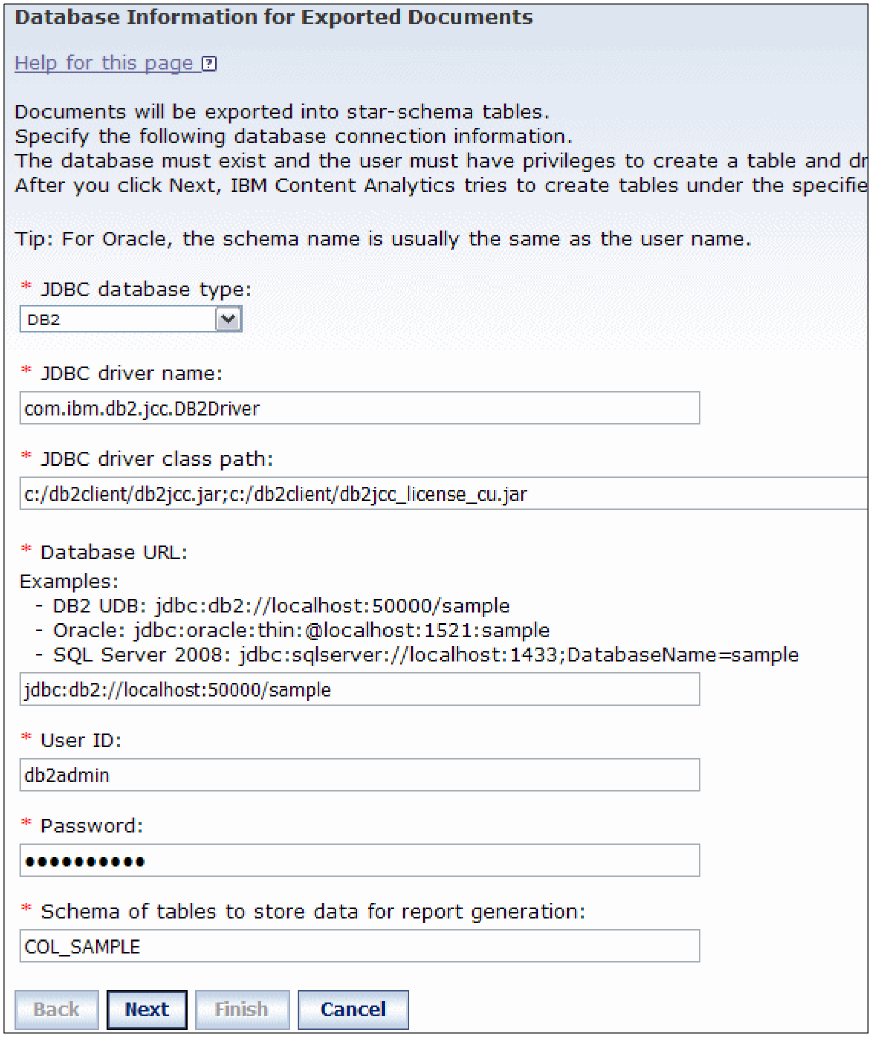
○ Export documents as CSV files

○ Export documents by using a custom plug-in

Figure 10-23 Export configuration into a relational database

c. In the Content and Fields to Export to a Database panel (Figure 10-24), enter values for the database URL, user name, password, and class path variables according to your environment. Then click **Next**.

> **Table and database creation:** Content Analytics automatically creates the tables you specified but you *must* create a database or use an existing database when editing the database URL.



Database Information for Exported Documents

Help for this page

Documents will be exported into star-schema tables.
Specify the following database connection information.
The database must exist and the user must have privileges to create a table and dr
After you click Next, IBM Content Analytics tries to create tables under the specifie

Tip: For Oracle, the schema name is usually the same as the user name.

* JDBC database type: DB2

* JDBC driver name: com.ibm.db2.jcc.DB2Driver

* JDBC driver class path: c:/db2client/db2jcc.jar;c:/db2client/db2jcc_license_cu.jar

* Database URL:
Examples:
- DB2 UDB: jdbc:db2://localhost:50000/sample
- Oracle: jdbc:oracle:thin:@localhost:1521:sample
- SQL Server 2008: jdbc:sqlserver://localhost:1433;DatabaseName=sample

jdbc:db2://localhost:50000/sample

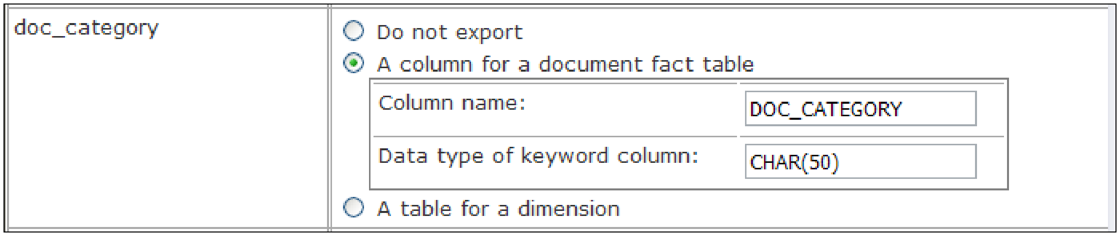
* User ID: db2admin

* Password: ●●●●●●●●●●

* Schema of tables to store data for report generation: COL_SAMPLE

Figure 10-24 Database Information for Exported Documents panel

5. In the Content and Fields to Export to a Database panel, select the fields to export. Set a desired column name, data type, and length for each exported field based on your data set. Figure 10-25 shows an example of setting a field to export. For this example, select the **A column for a document fact table** radio button in the doc_category row. Then click **Next**.



doc_category

○ Do not export

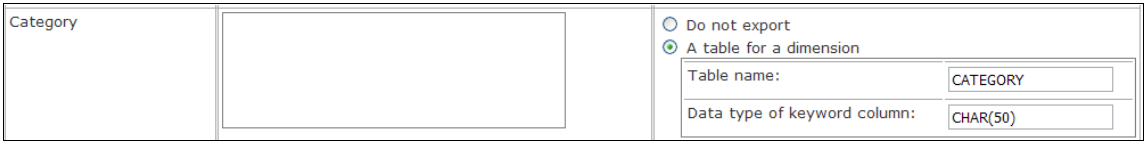
● A column for a document fact table

Column name:	DOC_CATEGORY
Data type of keyword column:	CHAR(50)

○ A table for a dimension

Figure 10-25 Setting a field to be exported

6. In the Facets to Export to a Database panel (Figure 10-26), select the facets that you want to export. Set a desired column name, data type, and length for each exported field based on your data set. Each selected facet to be exported will result in a new table being added to the database. In this scenario, select **A table for a dimension** in the Category row. Then click **Next**.



Category

○ Do not export

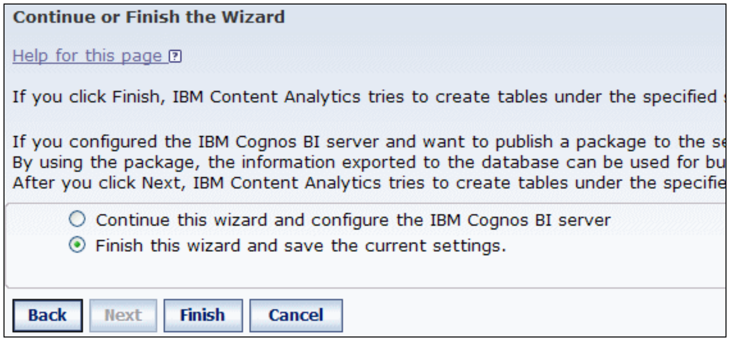
● A table for a dimension

Table name:	CATEGORY
Data type of keyword column:	CHAR(50)

Figure 10-26 Setting a facet to be exported to a database

7. In the Continue or Finish the Wizard panel (Figure 10-27), select the **Finish this wizard and save the current settings** radio button.

> **Further analysis:** You can perform further analysis by using data warehouse applications and generating reports such as Cognos Business Intelligence. To use the wizard to configure IBM Cognos BI server reports, select the **Continue this wizard and configure the IBM Cognos BI server** radio button (Figure 10-27). For further details, see Chapter 13, "Integrating Cognos Business Intelligence" on page 525.



Continue or Finish the Wizard

Help for this page

If you click Finish, IBM Content Analytics tries to create tables under the specified

If you configured the IBM Cognos BI server and want to publish a package to the se
By using the package, the information exported to the database can be used for bu
After you click Next, IBM Content Analytics tries to create tables under the specifie

○ Continue this wizard and configure the IBM Cognos BI server
● Finish this wizard and save the current settings.

Figure 10-27 Continue or Finish the Wizard window

Then click **Finish**.

As a result, Content Analytics attempts to create tables under the specified database.

Running the export

At this point, the export to a database is configured and the database tables are created. To export the analyzed documents, perform the following steps.

1. If your parse and index service is already started, restart the service.
2. If the collection does not have any crawled documents in the index, start the crawler, parse, and index components. If you have already crawled and parsed documents, also perform a full build of index.

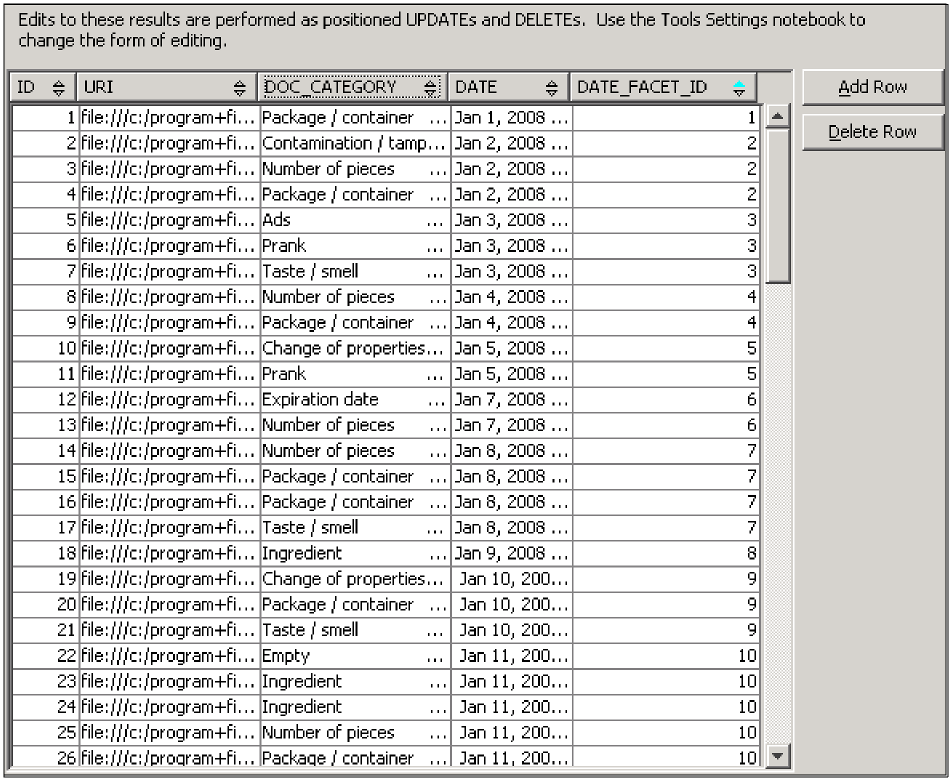
Validating the results of the analyzed documents export

Using Export Monitoring in the administration console, validate that the export request that has been queued by the Content Analytics server. Additionally, access the database, and check the number of records in the DOC_FACT table.

For example, if you used the default database table and schema names, enter the following command in the DB2 command window:

```
SELECT COUNT(*) FROM COL_SAMPLE.DOC_FACT
```

The returned number is the same as the number of documents that were analyzed. Figure 10-28 shows the result of the exported data in a DB2 database.



Edits to these results are performed as positioned UPDATEs and DELETEs. Use the Tools Settings notebook to change the form of editing.

ID	URI	DOC_CATEGORY	DATE	DATE_FACET_ID
1	file:///c:/program+fi…	Package / container …	Jan 1, 2008 …	1
2	file:///c:/program+fi…	Contamination / tamp…	Jan 2, 2008 …	2
3	file:///c:/program+fi…	Number of pieces …	Jan 2, 2008 …	2
4	file:///c:/program+fi…	Package / container …	Jan 2, 2008 …	2
5	file:///c:/program+fi…	Ads …	Jan 3, 2008 …	3
6	file:///c:/program+fi…	Prank …	Jan 3, 2008 …	3
7	file:///c:/program+fi…	Taste / smell …	Jan 3, 2008 …	3
8	file:///c:/program+fi…	Number of pieces …	Jan 4, 2008 …	4
9	file:///c:/program+fi…	Package / container …	Jan 4, 2008 …	4
10	file:///c:/program+fi…	Change of properties…	Jan 5, 2008 …	5
11	file:///c:/program+fi…	Prank …	Jan 5, 2008 …	5
12	file:///c:/program+fi…	Expiration date …	Jan 7, 2008 …	6
13	file:///c:/program+fi…	Number of pieces …	Jan 7, 2008 …	6
14	file:///c:/program+fi…	Number of pieces …	Jan 8, 2008 …	7
15	file:///c:/program+fi…	Package / container …	Jan 8, 2008 …	7
16	file:///c:/program+fi…	Package / container …	Jan 8, 2008 …	7
17	file:///c:/program+fi…	Taste / smell …	Jan 8, 2008 …	7
18	file:///c:/program+fi…	Ingredient …	Jan 9, 2008 …	8
19	file:///c:/program+fi…	Change of properties…	Jan 10, 200…	9
20	file:///c:/program+fi…	Package / container …	Jan 10, 200…	9
21	file:///c:/program+fi…	Taste / smell …	Jan 10, 200…	9
22	file:///c:/program+fi…	Empty …	Jan 11, 200…	10
23	file:///c:/program+fi…	Ingredient …	Jan 11, 200…	10
24	file:///c:/program+fi…	Ingredient …	Jan 11, 200…	10
25	file:///c:/program+fi…	Number of pieces …	Jan 11, 200…	10
26	file:///c:/program+fi…	Package / container …	Jan 11, 200…	10

Figure 10-28 DOC_FACT table in DB2 with exported analytics data

10.6.3 Exporting search result documents to the file system for Classification Module

With Content Analytics, you can export search results (documents) from the text miner application. You can export a limited set of documents for further analysis, monitoring, and reporting.

As shown in Figure 10-29, you can export the following items:

- Crawled content and metadata

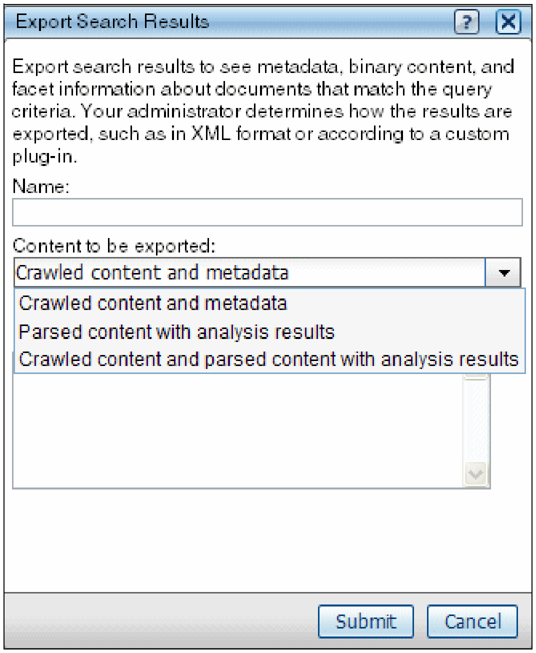
  Exporting with this option yields similar output as exporting crawled documents. With this option, Content Analytics exports the native content and metadata as explained in 10.2.1, “Crawled documents” on page 397, for the search results. The content is exported as a `.dat` file or as native content.

- Parsed content with analysis results

  When you export search results with this option, Content Analytics exports metadata, facets, and extracted text. Facets and extracted text are included in the metadata file.

- Crawled content and parsed content with analysis results

  With this option, the exported output is a combination of the first two options.



Export Search Results

Export search results to see metadata, binary content, and facet information about documents that match the query criteria. Your administrator determines how the results are exported, such as in XML format or according to a custom plug-in.

Name:

Content to be exported: Crawled content and metadata

Crawled content and metadata
Parsed content with analysis results
Crawled content and parsed content with analysis results
Submit
Cancel

Figure 10-29 Export options for the search result documents

To export search results to the file system for Classification Module, use the following procedure:

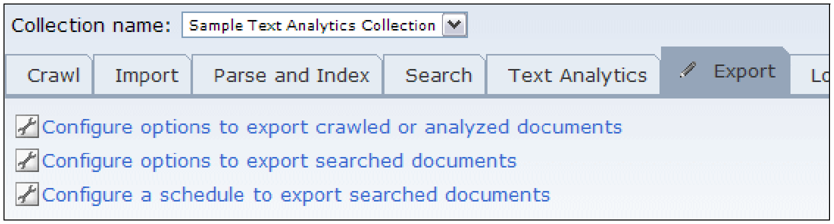
1. Configure export for the search results by using the administration console.
2. Perform a search, and export the search results in the text miner application.

You can optionally schedule an export for a later time by using the administration console after the request is made. Use the steps in the following section to get started.

Configuring export for the search results

To configure the export option for the search results in Content Analytics, follow these steps:

1. From the administration console, click **Collections** from the toolbar. Locate the collection that you want to edit, and click **Edit**.
2. Select the **Export** tab (Figure 10-30), and click the **Configure options to export searched documents** link.



Collection name: Sample Text Analytics Collection

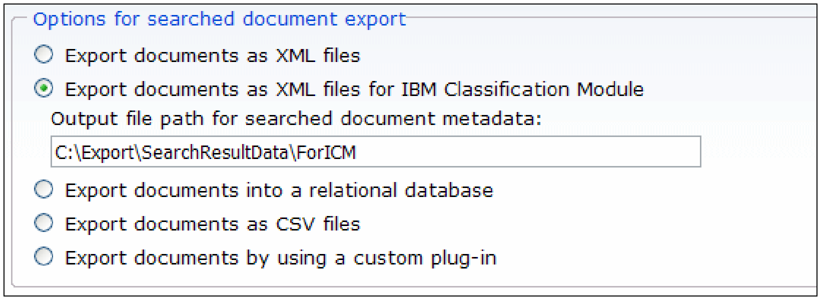
Configure options to export crawled or analyzed documents

Configure options to export searched documents

Configure a schedule to export searched documents

Figure 10-30 Export tab

3. Under Options for searched document export (Figure 10-31), select **Export documents as XML files for InfoSphere Classification Module**, and enter the path of an existing directory for the output.



Options for searched document export

○ Export documents as XML files

● Export documents as XML files for IBM Classification Module

Output file path for searched document metadata:

C:\Export\SearchResultData\ForICM

○ Export documents into a relational database

○ Export documents as CSV files

○ Export documents by using a custom plug-in

Figure 10-31 Export search result documents

4. Click **OK**.

Performing a search and exporting the search result

After you configure the export for the search results, perform a search and export the search result. For illustration purposes, this example shows the steps to search for complaints about ice cream.

To search and export the search result, follow these steps:

1. Launch the text miner application, and click the **Expand this area** icon at the top of the panel to view the query text area.
2. Enter `ice cream` as the search term in the text area, and click **Search** (Figure 10-32).

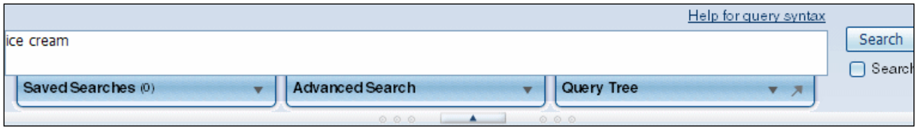
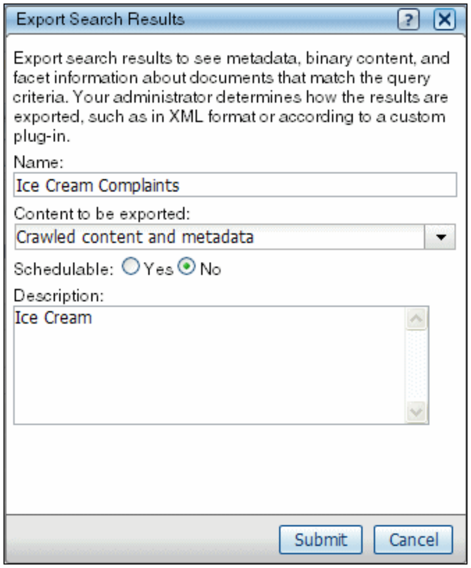


Figure 10-32 Executing the search

3. Click the **Export** icon.
4. In the Export Search Results window (Figure 10-33), complete the following tasks to configure the export:
   a. In the Name field, enter `Ice Cream Complaints`.
   b. For Content to be exported, enter `Crawled content and metadata`.
   c. For Schedulable, click **No** or **Yes**. If you select **Yes**, you must configure a schedule in the administration console. Select **No**.
   d. Enter a description of `Ice Cream`.
   e. Click **Submit**.



Export Search Results

Export search results to see metadata, binary content, and facet information about documents that match the query criteria. Your administrator determines how the results are exported, such as in XML format or according to a custom plug-in.

Name: Ice Cream Complaints

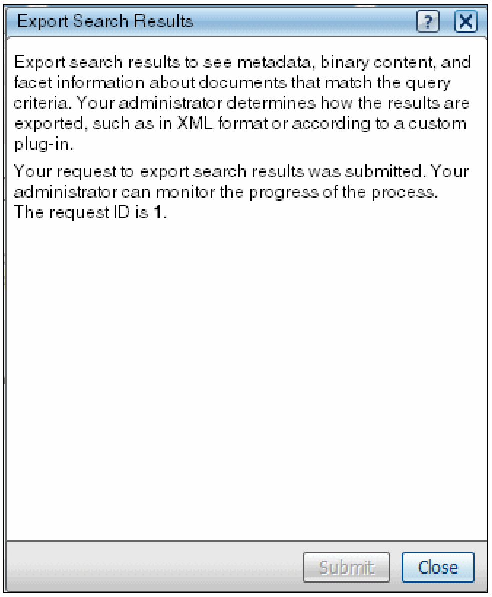
Content to be exported: Crawled content and metadata

Schedulable: ○ Yes ● No

Description: Ice Cream

Figure 10-33 Export Search Results window

5. In the confirmation window (Figure 10-34), click **Close**.



Export search results to see metadata, binary content, and facet information about documents that match the query criteria. Your administrator determines how the results are exported, such as in XML format or according to a custom plug-in.

Your request to export search results was submitted. Your administrator can monitor the progress of the process. The request ID is **1**.

Figure 10-34 Export Search Results showing confirmation with a request ID

If you configured a scheduled export, continue to the next section. Otherwise, jump to "Validating the export" on page 426.

Optional: Scheduling an export

Scheduling an export is an optional feature of Content Analytics. To schedule for an export, follow these steps:

1. Perform a search in the text miner application following steps similar to the steps in "Performing a search and exporting the search result" on page 422, except select **Yes** for the Schedulable radio button.
2. From the administration console, click **Collections** in the toolbar. Locate the collection that you want to edit, and click **Edit**.
3. Click the **Export** tab (Figure 10-35) and click the **Configure a schedule to export searched documents** link.

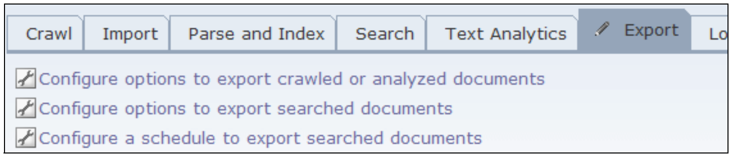

Crawl
Import
Parse and Index
Search
Text Analytics
Export
Lo
Configure options to export crawled or analyzed documents
Configure options to export searched documents
Configure a schedule to export searched documents

Figure 10-35 Scheduling an export for searched documents

4. Under Specify a general schedule (Figure 10-36), follow these steps:

   a. For the Start on field, select the appropriate values for hours, minutes, and time zone.

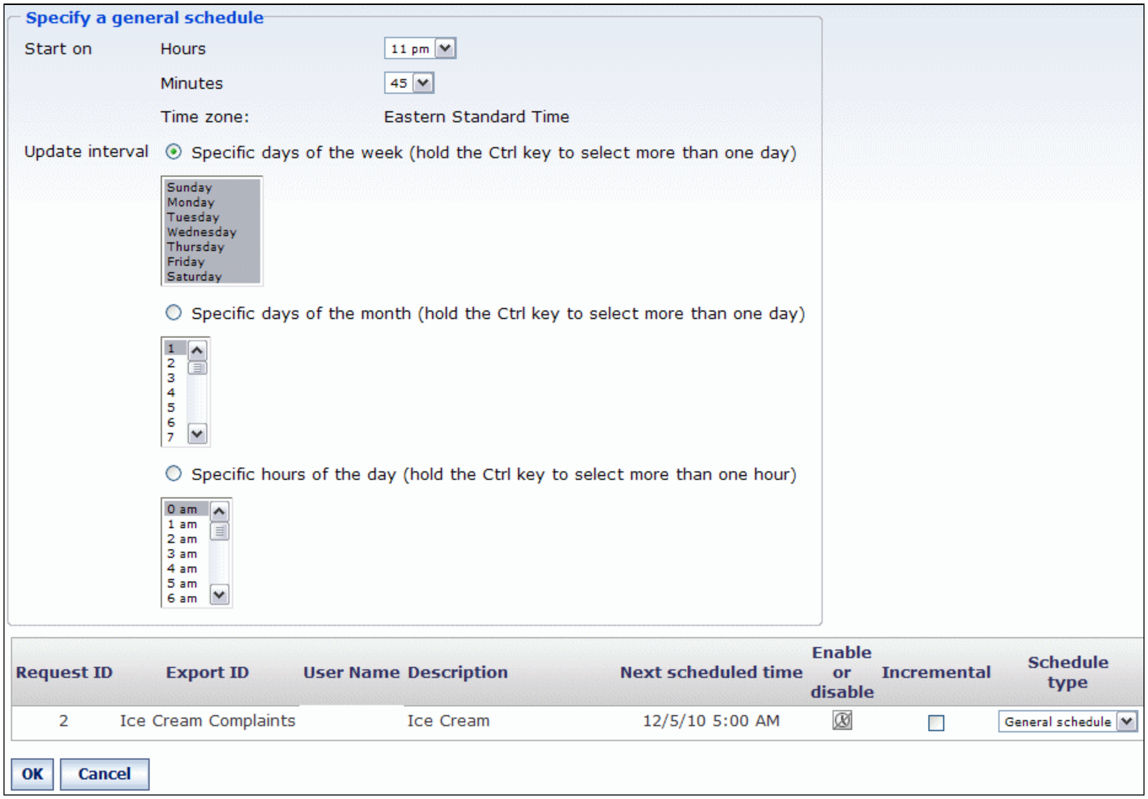
   b. For the Update interval field, select the appropriate values for the specific days of the week, month, and hours.

   c. In the bottom table, locate your request id for your exported scheduled search. Click the **Enable** icon to enable the export schedule.

   d. Optional: In the Schedule Type field, select **Custom**, and specify a custom schedule. You specify the custom schedule by clicking the **Configure** icon next to the schedule type field.

   e. Optional: Select the **Incremental Export** check box.

   f. Click **OK**.



Specify a general schedule

Start on
Hours: 11 pm
Minutes: 45
Time zone: Eastern Standard Time

Update interval
● Specific days of the week (hold the Ctrl key to select more than one day)
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

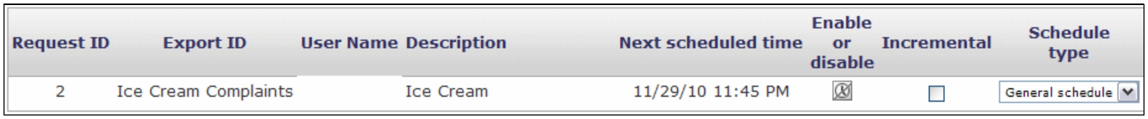
○ Specific days of the month (hold the Ctrl key to select more than one day)
1
2
3
4
5
6
7

○ Specific hours of the day (hold the Ctrl key to select more than one hour)
0 am
1 am
2 am
3 am
4 am
5 am
6 am

Request ID	Export ID	User Name	Description	Next scheduled time	Enable or disable	Incremental	Schedule type
2	Ice Cream Complaints		Ice Cream	12/5/10 5:00 AM		□	General schedule

Figure 10-36 Configuring the schedule to export the search results

5. Click the **Configure a schedule to export search documents** link. In the Next scheduled time field (Figure 10-37), a time is set for the next scheduled export.



Request ID	Export ID	User Name	Description	Next scheduled time	Enable or disable	Incremental	Schedule type
2	Ice Cream Complaints		Ice Cream	11/29/10 11:45 PM	☒	□	General schedule ▼

Figure 10-37 Scheduled export of searching the result documents

Validating the export

Using Export Monitoring in the administration console, validate that the export request has been queued by the Content Analytics server. Check the output directory that you specified to verify that data is being exported.

For the request that we submitted, two files are created in the `C:\Export\SearchResultData\ForICM` directory:

- `catalog.xml`
- `Ice Cream Complaints.xml`

Example 10-5 shows a partial `catalog.xml` file.

Example 10-5 Partial catalog.xml file

```
<?xml version="1.0" encoding="UTF-8"?>
<_Catalog entry_count="19">
  <Entry display_name="Body" type="string" nlp_usage="PlainText" is_viewed="true"
is_categories="false" is_link="false" is_matches="false" is_scores="false"
is_firedRules="false" is_changedNVPs="false"><![CDATA[Body]]></Entry>
......
</_Catalog>
```

Example 10-6 shows a partial `Ice Cream Complaints.xml` file.

Example 10-6 Partial Ice Cream Complaints.xml file

```
<?xml version="1.0" encoding="UTF-8"?>
<Corpus_Bundle>
  <Corpus_Item>
    <ICM_NVP key="Body">
00000850
vanilla ice cream - Taste / smell
2008-12-30
1230635992343
Taste / smell
```

```
Strange odor
vanilla ice cream
I bought some ice cream today, but it had a strange odor.</ICM_NVP>
    <ICM_NVP key="date">1230635992343</ICM_NVP>
    <ICM_NVP
key="directory">C:&#x5C;IBM&#x5C;es&#x5C;samples&#x5C;firststep&#x5C;data&#x5C;xml&#x
5C;xml-data</ICM_NVP>
    <ICM_NVP key="doc_category">Taste / smell</ICM_NVP>
    <ICM_NVP key="doc_id">00000850</ICM_NVP>
    <ICM_NVP key="doc_product">vanilla ice cream</ICM_NVP>
    <ICM_NVP key="doc_subcategory">Strange odor</ICM_NVP>
    <ICM_NVP
key="docid">file:///C:/IBM/es/samples/firststep/data/xml/xml-data/00000850.xml</ICM_N
VP>
    <ICM_NVP key="extension">.xml</ICM_NVP>
    <ICM_NVP key="filename">00000850.xml</ICM_NVP>
    <ICM_NVP key="filesize">395</ICM_NVP>
    <ICM_NVP key="modifieddate">1250055768000</ICM_NVP>
    <ICM_NVP key="title">vanilla ice cream - Taste / smell</ICM_NVP>
    <ICM_NVP key="Category">Ice Cream</ICM_NVP>
  </Corpus_Item>
  <Corpus_Item>
......
</Corpus_Item>
</Corpus_Bundle>
```

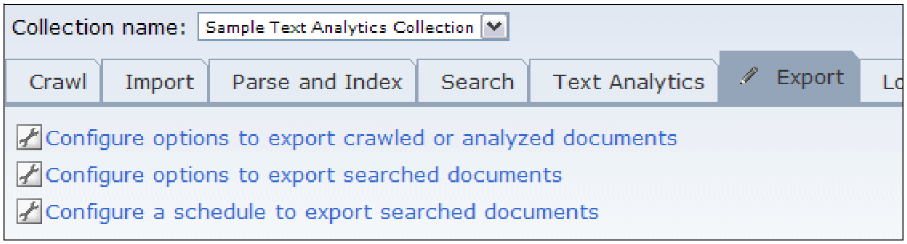
10.6.4 Exporting search result documents to CSV files

With Content Analytics, you can export crawled documents, analyzed documents, or search results to CSV files on the file system. By following this approach, you can work with the data outside of Content Analytics. The CSV files conform to the RFC 4180 standard. The files are delimited by a comma to identify each column. Moreover, the generated CSV files are imitated star-schema tables that are similar to the star-schema created during the document export to relational database functionality.

Configuring the search export to CSV files

To configure the export to CSV files in Content Analytics, follow these steps:

1. From the administration console, click **Collections** from the toolbar. Locate the collection that you want to edit, and click **Edit**.
2. Select the **Export** tab (Figure 10-38) and click the **Configure options to export searched documents** link.



Collection name: Sample Text Analytics Collection

Crawl | Import | Parse and Index | Search | Text Analytics | Export | Lo

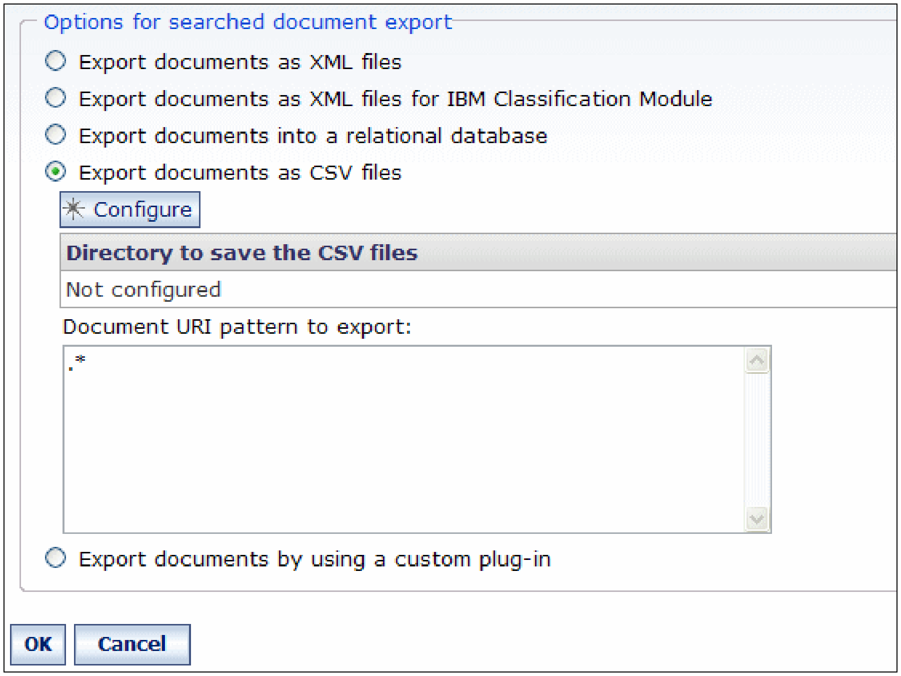
Configure options to export crawled or analyzed documents

Configure options to export searched documents

Configure a schedule to export searched documents

Figure 10-38 Export tab

3. Under Options for searched document export (Figure 10-39), select **Export documents as CSV Files** and click **Configure**.



Options for searched document export

○ Export documents as XML files

○ Export documents as XML files for IBM Classification Module

○ Export documents into a relational database

● Export documents as CSV files

Configure

Directory to save the CSV files

Not configured

Document URI pattern to export:

```
.*
```

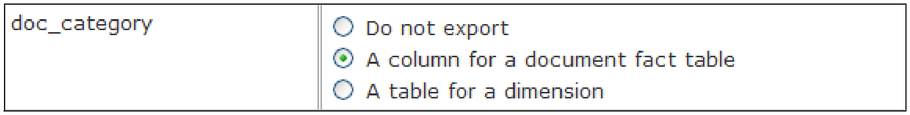
○ Export documents by using a custom plug-in

OK Cancel

Figure 10-39 Export documents as CSV files option

4. Select the search fields that you want to export. You can export the field to a column in the document fact table by selecting the **A column for a document fact table** radio option associated to the field. Instead, you can select the **A table for a dimension** radio option to export the search field to its own CSV file that can be imported into a relational database as an individual table.

   In this scenario, select the **A column for a document facet table** radio button for the doc_category field, as shown in Figure 10-40. Then click **Nex**t.

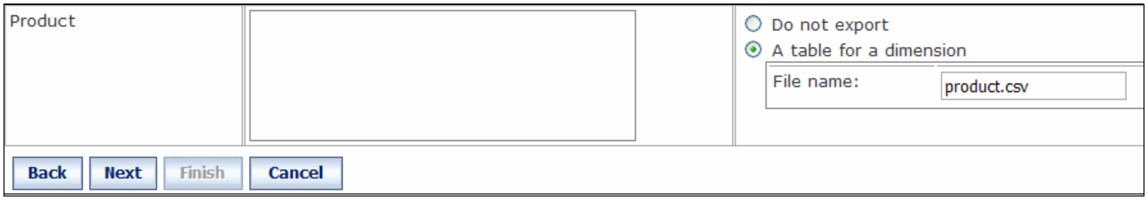


doc_category

○ Do not export
● A column for a document fact table
○ A table for a dimension

Figure 10-40 Fields to export to CSV files

5. Select the facets that you want to export. You can select the **A table for a dimension** radio option to export the search field to its own CSV file so that it can be imported into a relational database as an individual table. For this scenario, select the **A table for a dimension** radio option for the **Product** facet, as shown in Figure 10-41. Then click **Next**.



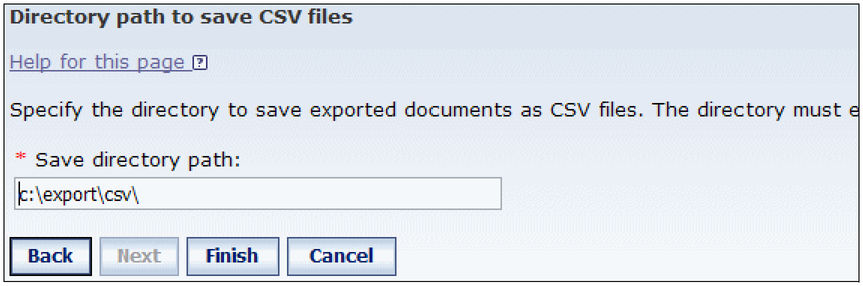
Product

○ Do not export
● A table for a dimension

File name: product.csv

Figure 10-41 Facet to export to CSV files

6. In the Directory path to save CSV files panel (Figure 10-42), enter the directory path where the exported documents will be exported. For this scenario, type `c:\export\csv` in the Save directory path field. Create the directory path, `c:\export\csv`, on your machine because it must exist to configure the export. Then click **Finish**.



Directory path to save CSV files

Help for this page

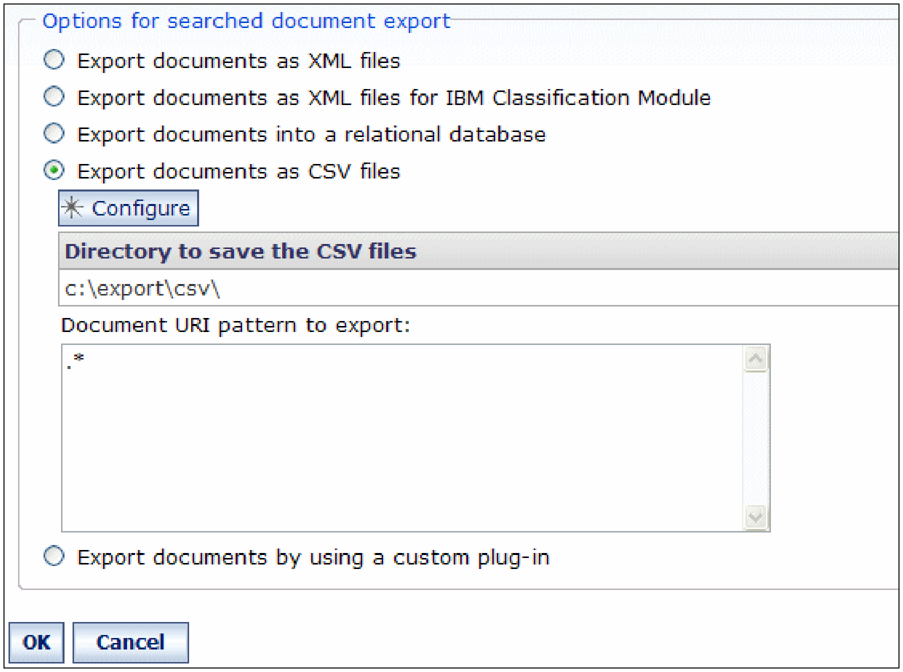
Specify the directory to save exported documents as CSV files. The directory must e

* Save directory path:

c:\export\csv\

Figure 10-42 Directory path to save the CSV files

7. Under Options for searched document export (Figure 10-43), if you want to limit the type of document that will be exported, add the document type to the Document URI pattern to export field. In this scenario, keep the default value of **.***, which exports all document types. Then click **OK**.



Options for searched document export

○ Export documents as XML files

○ Export documents as XML files for IBM Classification Module

○ Export documents into a relational database

● Export documents as CSV files

Directory to save the CSV files: c:\export\csv\

Document URI pattern to export: .*

○ Export documents by using a custom plug-in

Figure 10-43 Search document export options

Exporting search files to CSV files

After you configure the export to CSV files, perform a search, and export the search results to CSV files. Follow the instructions in "Performing a search and exporting the search result" on page 422. After you perform the search and export steps, the crawled content and metadata for documents that contain the term ice cream are exported to CSV files.

Validating exported CSV files

Using Export Monitoring in the administration console, validate that the export request has been queued by the Content Analytics server. Check the output directory that you specified to verify that data is being exported.

For the request that we submitted, the following files are created in the `C:\export\csv` directory:

date_facet.csv Contains the ID and Keyword fields as columns. Our document does not have any exported dates. Therefore, this file is empty.

doc_fact.csv Contains the ID, URI, doc_category (search field that we previously configured for export), DATE, and DATE_FACET_ID.

doc_flag.csv Contains the ID and Name for each document flag. Our example doe not contain any exported document flags. Therefore, this file is empty.

doc_flag_brg.csv Provides data to link the document flag with the specific document. It contains the ID for the file that matches the ID column value in the `doc_fact.csv` file and the ID of the document flag listed in the `doc_flag.csv` file.

export_fact.csv Contains the document ID of the exported documents.

export_metadata.csv Provides data related to the export. It contains the REQID, EXPORTID, DESCRIPTION, QUERY, and USER for the export request.

product.csv Contains the ID and Product facet data.

product_brg.csv Provides data to link the product facet with the specific document. It contains the ID for the file that matches the ID column value in the `doc_fact.csv` file and the ID of the product facet listed in the `product.csv` file.

10.7 Deep inspection

With the deep inspection feature of Content Analytics, you can export the entire analysis of a set of documents that match a query defined in the text miner application. The analysis and statistics generated by deep inspection are the same as those performed by the text miner application but without limits imposed on the number of facets and their values.

The text miner application is designed as an adhoc text mining tool that supports rapid calculation in response to frequent changes in your query. If the analyzed data contains a large number of facets and keywords, performance might be degraded. Consequently, limits are placed on the number of keywords (less than 500) that can be processed by the text miner application to prevent this kind of performance degradation.

In most cases, the limit does not affect your analysis. Discovery of trends, patterns, and high correlations usually surfaces in the most frequently occurring documents of your first 500 keywords (depending on your sort criteria). However, it is possible that, in certain scenarios, you want to view the entire set of calculations for all documents. You do that with deep inspection.

You enable the deep inspection function from the administration console. After enabling it, a deep inspection icon is enabled on each of the text miner views, except for the Documents view. When you click the **Deep Inspection** icon (), a batch submission is made for subsequent background processing of your query. The results of the deep inspection are stored in an XML file on the file system. The exported results from the deep inspection contain the selected keywords along with their frequency counts and correlation values. Deep inspection provides the same results as viewed from the text miner application but with deeper analysis.

You can schedule a deep inspection analysis and obtain reports on a periodic basis. Optionally, you can also submit the deep inspection request at any single point to view the analysis at an unscheduled time. The reports are exported as XML files. For details about content and format of the XML files, go to the IBM Content Analytics Information Center at the following address, and search on *XML file format for deep inspection reports*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

The deep inspection feature is often confused with exporting search results because both features allow exporting the data of interest at any given time in the text miner application. When you export search results, you get metadata, binary content, and facet information about the individual documents that match the query criteria. The exported content is about individual documents. With the deep inspection feature, you get *analysis statistics*, such as the top keywords, trends, deviation, and correlation values, on the documents. With such statistics, a business analyst can create a custom application to compare deep inspection reports to detect any anomalies.

For example, a car manufacturing company that is analyzing incident reports can set up a query in the text miner application to find incidents on battery failure. The company can also run a deep inspection report based on the frequency of incidents on a weekly basis. By running the deep inspection reports on a weekly basis, the manufacturer determines that, on average, they encounter 10 battery failure incidents per week. However, the frequency of battery failure increased drastically to 20 for the last week and 17 the week before. This type of anomaly is detected by the trend and deviation analysis and can serve as an alert to the business analyst.

This section provides information about the format and location of the exported deep inspection data followed by common configuration options. It also guides you through configuring and running a deep inspection on a sample collection created in Chapter 4, “Installing and configuring IBM Content Analytics” on page 71.

To obtain deep inspection reports, follow these steps:

1. Enable the deep inspection feature by using the administration console.
2. Issue deep inspection requests for analysis on a set of data.

You can monitor the request and validate the reports that are generated.

10.7.1 Location and format of the exported data

The location of the deep inspection report is similar to the location for exported documents as explained in 10.3.1, “Location of the exported data” on page 400. When you enable the deep inspection feature, you provide the path where the analysis statistics are to be exported. The output path of the analysis statistics generated by the deep inspection request depends on the name of the request and the time the request runs.

For example, if you create a deep inspection as XML request named `ProductFacetReport` on `March 30, 2010 at 5:25pm`, a new directory is created under the path. The name of the directory is based on the date and time at which the export occurs. The name of the created directory is in the *yyyymmddhhmm* format. The output files are created directly under the new directory. The report itself is in the XML format. Two files are created: one XML file that contains the analysis statistics and one XML file that contains information about the request.

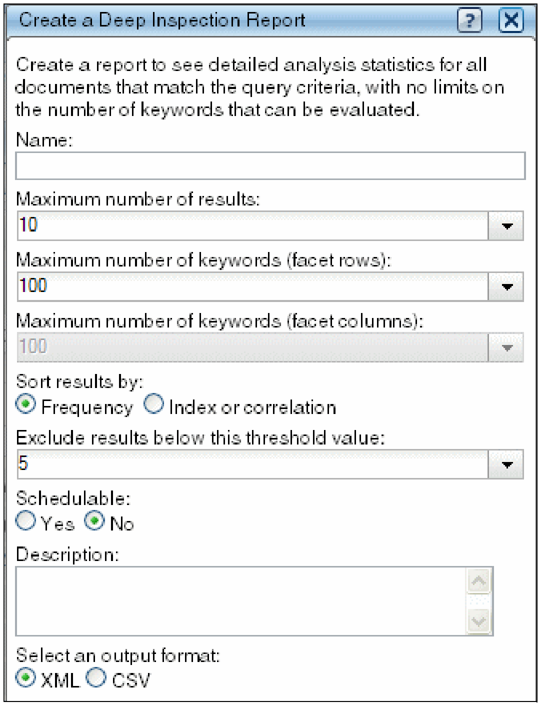
The example has the following report output:

```
C:\Export\DeepInspectionReports\201003301725\
ProductFacetReport.xml
ProductFacetReport_log.xml
```

For an example of the generated reports, see 10.7.7, “Validating the deep inspection reports generation” on page 445.

10.7.2 Common configuration

To obtain meaningful analysis statistics from the deep inspection feature, you must understand the different configuration options and how they influence the analysis result. Figure 10-44 shows the configuration options when submitting a deep inspection request.



Create a Deep Inspection Report

Create a report to see detailed analysis statistics for all documents that match the query criteria, with no limits on the number of keywords that can be evaluated.

Name:

Maximum number of results: 10

Maximum number of keywords (facet rows): 100

Maximum number of keywords (facet columns): 100

Sort results by: ● Frequency ○ Index or correlation

Exclude results below this threshold value: 5

Schedulable: ○ Yes ● No

Description:

Select an output format: ● XML ○ CSV

Figure 10-44 Deep inspection request configuration

Set the following options for a deep inspection request:

- Name

  This field refers to the name for the deep inspection report. You can enter your own name or accept the default value.

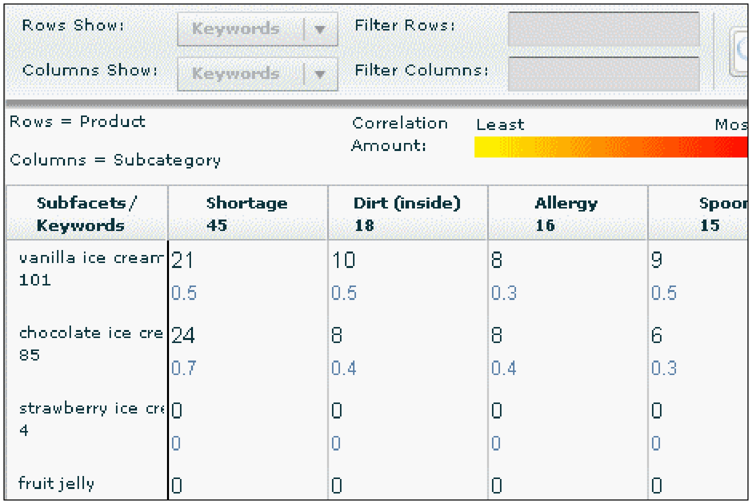
- Maximum number of results

  This field refers to the maximum number of analysis results to include in the generated report. You can select a value from the drop-down menu or enter your own value. For example, you can select the top 10 facets or the top 50 most correlated pairs of keywords for the two selected facets.

- Maximum number of keywords (facet rows)

  This option refers to the maximum number of the most frequent keywords to include when calculating analysis statistics result. For example, if the maximum number of keywords is set to 1000 and the maximum number of results is set to 10, the deep inspection report contains the top ten most frequent or correlated result out of top 1000 results.

  Figure 10-45 shows an example of Facet Pair view in the text miner application. In this view, the keywords “vanilla ice cream,” “chocolate ice cream,” “strawberry ice cream,” and “fruit jelly” are top keywords for the Product facet that will be considered when calculating the frequency or correlation among the documents. If the maximum number of keywords is set to 3, only the facet values “vanilla ice cream,” “chocolate ice cream,” and “strawberry ice cream” will be used. The reason is that they are the top three most frequent keywords for the Product facet.



Rows = Product

Columns = Subcategory

Subfacets / Keywords	Shortage 45	Dirt (inside) 18	Allergy 16	Spoor 15
vanilla ice crear 101	21 0.5	10 0.5	8 0.3	9 0.5
chocolate ice cre 85	24 0.7	8 0.4	8 0.4	6 0.3
strawberry ice cre 4	0 0	0 0	0 0	0 0
fruit jelly	0	0	0	0

Figure 10-45 Facet pair view in table format

- Maximum number of keywords (facet columns)

  This option is enabled for the Facet Pairs view where the configuration for the second facet is necessary to perform facet pair analysis. When analyzing facet pairs, you also select the maximum number of most frequent keywords for the second facet. For example, in Figure 10-45, the values “Shortage,” “Dirt (inside),” and “Allergy” are the top keywords for the facet subcategory, and they are displayed in the maximum number of keywords as facet columns.

- Sort result by

  With this option, you can choose to sort the results by frequency or correlation.

  Deep inspection calculation: Even though you can specify sorting by index or correlation, Content Analytics always finds the most frequent keywords first for the selected facet or facet pair. Content Analytics then calculates the top correlation or frequency among the most frequent keywords or facet pair values.

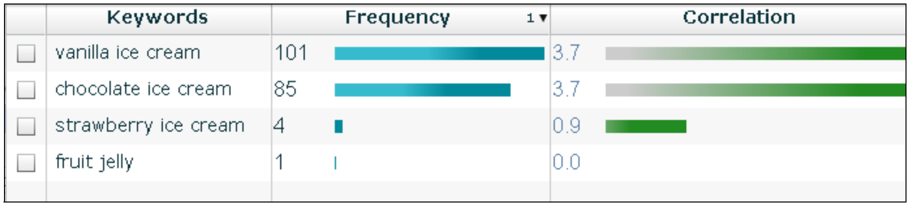
  For example, as shown in Figure 10-45 on page 435, you select Product as a facet row and Subcategory as a facet column in the Facet Pair view. When you submit a deep inspection request, you set 1000 for Maximum number of keywords (facet rows) field and 1000 for Maximum number of keywords (facet columns) field. In this case, Content Analytics selects the 1000 most frequently keywords for both the Product and Subcategory facets. It also computes a correlation and frequency (up to 1,000,000 correlation or frequency values). Depending on whether you select frequency or correlation for the Sort results by option, the list is generated and exported in the report.

  Lastly, the deep inspection report contains results based on the value set for the Maximum number of results. For example, if the value for Maximum number of results is set to 100, the deep inspection report contains 100 out of 1,000,000 correlation or frequency values.

  To find the most correlated data in less frequent keywords, you can select a larger number for Maximum number of keywords for facet rows and columns. By using this setting, deep inspection takes longer to generate the report.

- Exclude results below this threshold value

  This option refers to the cutoff threshold of the result set. For example, for the result in Figure 10-46, if you set a threshold of 80 and sort by frequency, only vanilla ice cream and chocolate ice cream are included in the report.



	Keywords	Frequency 1 ▾	Correlation
□	vanilla ice cream	101	3.7
□	chocolate ice cream	85	3.7
□	strawberry ice cream	4	0.9
□	fruit jelly	1	0.0

Figure 10-46 Sample analysis result

If you set the threshold value to 5 and sort by correlation, the deep inspection report does not contain any results because the highest correlation value of 3.7 is less than 5.

- Schedulable

  You can define the request as schedulable or as a one-time request. If you select the request as schedulable, see 10.7.5, “Optional: Scheduling a deep inspection run” on page 441, to configure a schedule. You must configure a schedule to obtain a deep inspection report.

- Description

  Use this field to specify a description of the request. The description can be any string value.

10.7.3 Enabling deep inspection

You enable the deep inspection feature from the administration console. The icon for invoking a deep inspection request is enabled in the text miner application *only* if you enable it first from the administration console for the collection that you work on.

To enable deep inspection, follow these steps:

3. From the administration console, click **Collections** in the toolbar.
4. In the Collections view, locate the collection that you want to edit, and click **Edit**.
5. Click the **Text Analytics** tab (Figure 10-47), and click the **Configure deep inspection options** link.

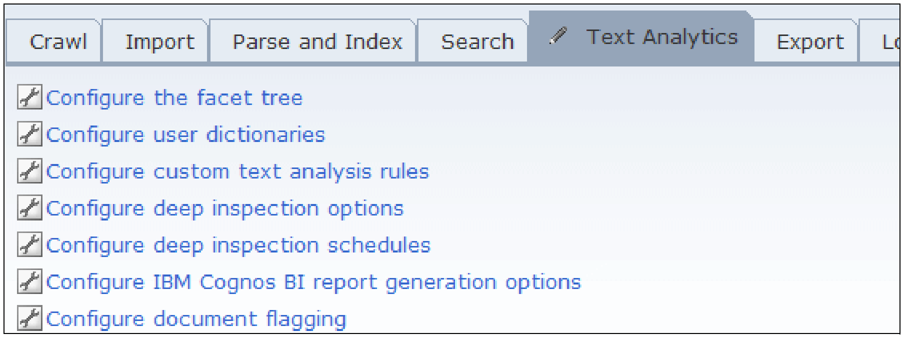
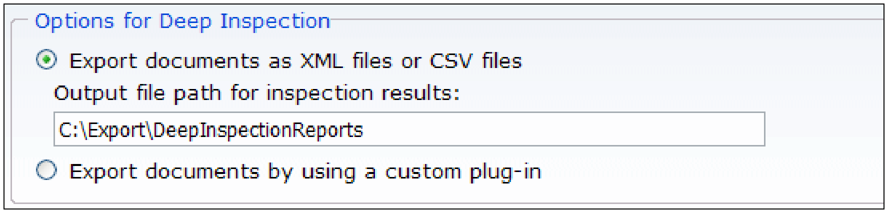

Crawl
Import
Parse and Index
Search
Text Analytics
Export
Lo
Configure the facet tree
Configure user dictionaries
Configure custom text analysis rules
Configure deep inspection options
Configure deep inspection schedules
Configure IBM Cognos BI report generation options
Configure document flagging

Figure 10-47 Selecting to configure the deep inspection options

6. Under Options for Deep Inspection (Figure 10-48), select **Export documents as XML files or CSV files**. In the Output file path for inspection results field, enter an existing directory name. Then click **OK**.



Options for Deep Inspection

● Export documents as XML files or CSV files

Output file path for inspection results:

C:\Export\DeepInspectionReports

○ Export documents by using a custom plug-in

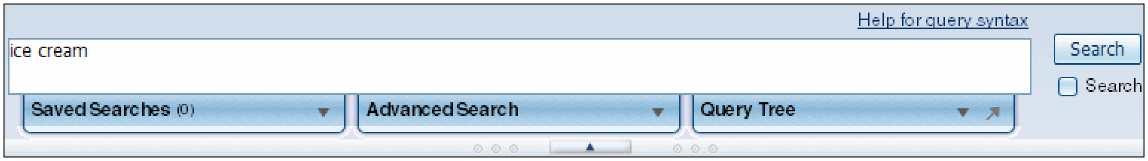
Figure 10-48 Configuring the options for deep inspection

10.7.4 Generating deep inspection reports

To generate a deep inspection report, submit a request using the text miner application. In this section, you are asked to click the **Deep Inspection** icon () to issue a deep inspection request.

To submit a request for a deep inspection report (on facets), follow these steps:

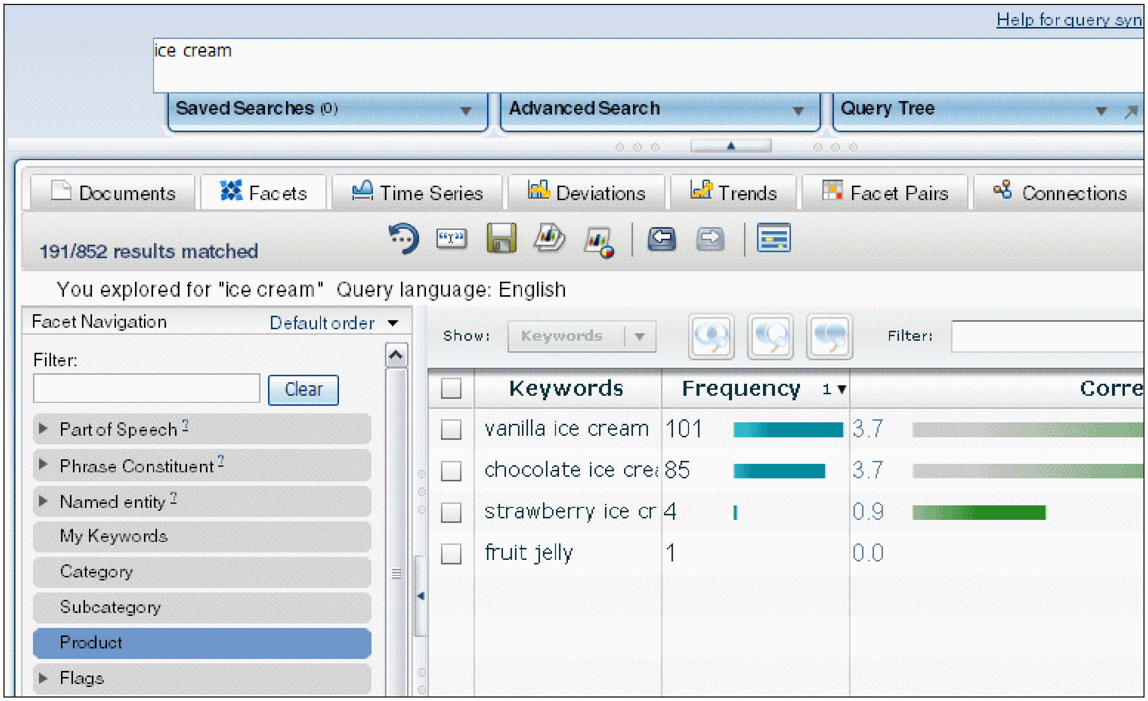
1. Access the text miner application.
2. Click the **Show query input area** link.
3. Enter search text (for example, `ice cream`), and click **Search** (Figure 10-49).



ice cream

Figure 10-49 Executing a search

4. Go to the Facets view, and click the **Product** facet. Four different flavors of ice creams are displayed (Figure 10-50).



ice cream

191/852 results matched

You explored for "ice cream" Query language: English

□	Keywords	Frequency 1 ▾	Corre
□	vanilla ice cream	101	3.7
□	chocolate ice cre:	85	3.7
□	strawberry ice cr	4	0.9
□	fruit jelly	1	0.0

Figure 10-50 Facet Pair view

5. Click the **Deep inspection** icon.

6. In the Create a Deep Inspection Report window (Figure 10-51 on page 440), complete the following steps:

   a. Enter a name.
   b. For Maximum number of results, select **10**.
   c. For Maximum number of keywords (facet rows), select **100**.
   d. For Sort results by, select **Frequency**.
   e. For Exclude results below this threshold value, select 5.
   f. For Schedulable, select **No**.

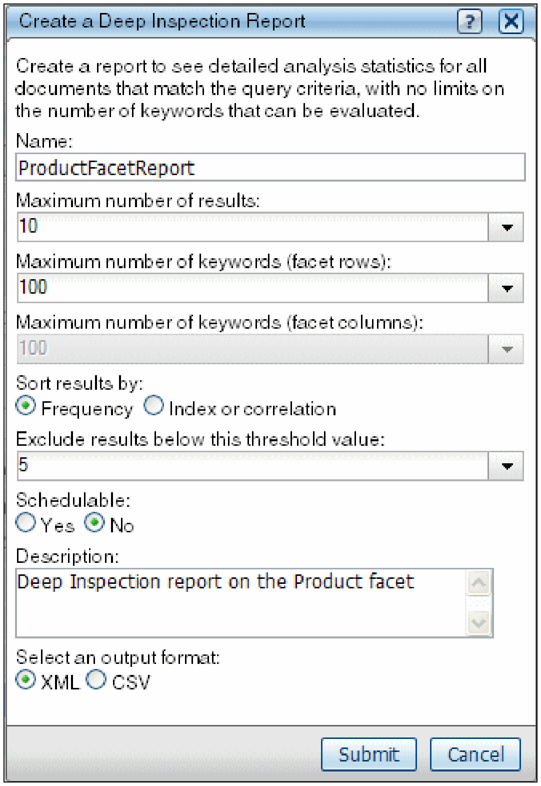
   > **Schedulable field:** If you select **Yes** for the Schedulable field, you must configure a schedule as explained in 10.7.5, “Optional: Scheduling a deep inspection run” on page 441.

   g. Enter a description.

h. For Select an output format, select **XML**.

> **Select an output format field:** You can export the deep inspection results as XML or CSV files. If you select the **CSV** option, the deep inspection report will be displayed in the CSV file format.

i. Click **Submit**.



Create a Deep Inspection Report

Create a report to see detailed analysis statistics for all documents that match the query criteria, with no limits on the number of keywords that can be evaluated.

Name: ProductFacetReport

Maximum number of results: 10

Maximum number of keywords (facet rows): 100

Maximum number of keywords (facet columns): 100

Sort results by: ● Frequency ○ Index or correlation

Exclude results below this threshold value: 5

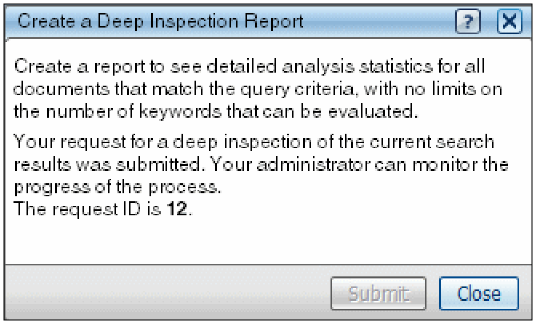
Schedulable: ○ Yes ● No

Description: Deep Inspection report on the Product facet

Select an output format: ● XML ○ CSV

Figure 10-51 Submitting a deep inspection report request

7. In the Export Search Results confirmation window (Figure 10-52), click **Close**.



Create a Deep Inspection Report

Create a report to see detailed analysis statistics for all documents that match the query criteria, with no limits on the number of keywords that can be evaluated.

Your request for a deep inspection of the current search results was submitted. Your administrator can monitor the progress of the process.
The request ID is **12**.

Figure 10-52 Deep inspection request submitted

If you configured a scheduled export, continue to the next section. Otherwise, jump to 10.7.6, "Monitoring the deep inspection requests" on page 444.

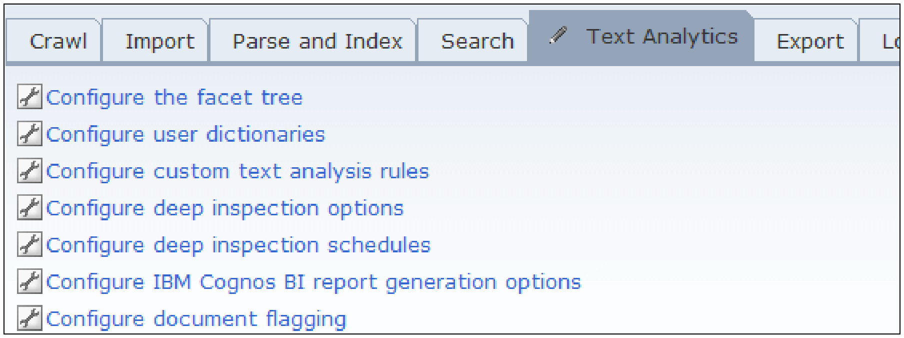
10.7.5 Optional: Scheduling a deep inspection run

With Content Analytics, you can schedule deep inspection analysis. The scheduling capability is similar to the capability for exporting the search results mentioned in 10.4.5, "Scheduling" on page 406, except for incremental export. That is, you cannot run deep inspection analysis on an incremental basis.

To schedule for a deep inspection run, follow these steps:

1. Generate a deep inspection report within the text miner application by following steps similar to the steps in 10.7.4, "Generating deep inspection reports" on page 438. The difference is that you must select **Yes** for the Schedulable radio button.
2. From the administration console, click **Collections** in the toolbar.
3. In the Collections view, locate the collection that you want to edit, and click **Edit**.

4. Click the **Text Analytics** tab (Figure 10-53), and click the **Configure deep inspection options** link.



Configure the facet tree
Configure user dictionaries
Configure custom text analysis rules
Configure deep inspection options
Configure deep inspection schedules
Configure IBM Cognos BI report generation options
Configure document flagging

Figure 10-53   Selecting the Configure deep inspection schedules link

5. Under Specify a general schedule (Figure 10-54 on page 443), follow these steps:

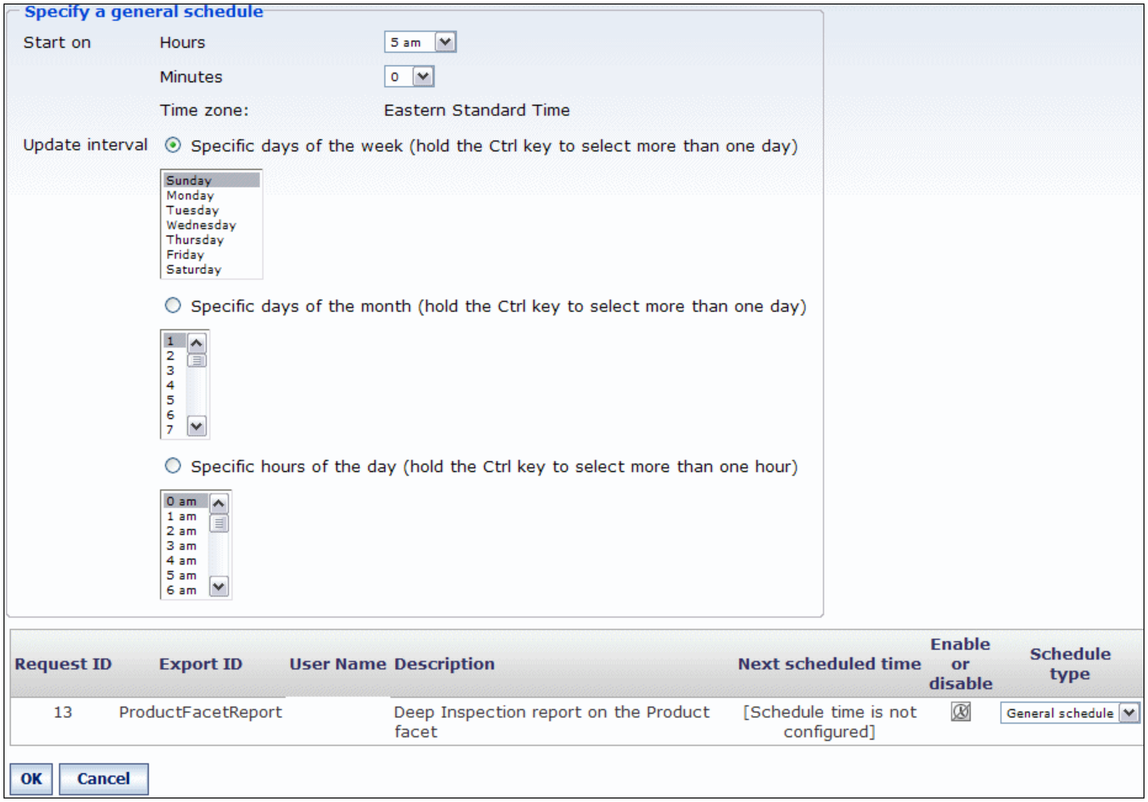
   a. For the Start on field, select the appropriate values for hours, minutes, and time zone.

   b. For the Update interval field, select the appropriate values for specific days of the week, month, and hours of the days.

c. Click the **Enable** icon to enable the export, as shown in Figure 10-54.

d. Optional: for the Schedule Type field, select **Custom**, and specify a custom schedule.

e. Click **OK**.



Specify a general schedule

Start on Hours: 5 am

Minutes: 0

Time zone: Eastern Standard Time

Update interval ● Specific days of the week (hold the Ctrl key to select more than one day)

Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

○ Specific days of the month (hold the Ctrl key to select more than one day)

1
2
3
4
5
6
7

○ Specific hours of the day (hold the Ctrl key to select more than one hour)

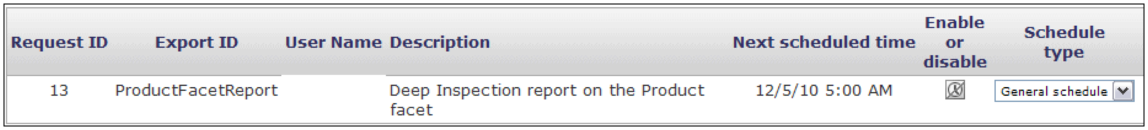
0 am
1 am
2 am
3 am
4 am
5 am
6 am

Request ID	Export ID	User Name	Description	Next scheduled time	Enable or disable	Schedule type
13	ProductFacetReport		Deep Inspection report on the Product facet	[Schedule time is not configured]	☒	General schedule

OK Cancel

Figure 10-54 Configure schedule for deep inspection request

6. Click the **Configure deep inspection schedules** link. You see a time for the next scheduled run for deep inspection (Figure 10-55).



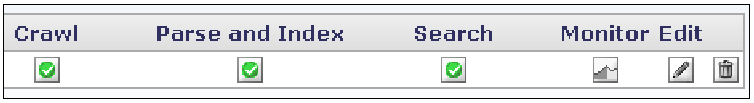
Request ID	Export ID	User Name	Description	Next scheduled time	Enable or disable	Schedule type
13	ProductFacetReport		Deep Inspection report on the Product facet	12/5/10 5:00 AM	☒	General schedule

Figure 10-55 Scheduled deep inspection request

10.7.6 Monitoring the deep inspection requests

Content Analytics provides a monitoring capability from the administration console to help you see the status of deep inspection analysis requests. You can validate that requests are created by using the following steps:

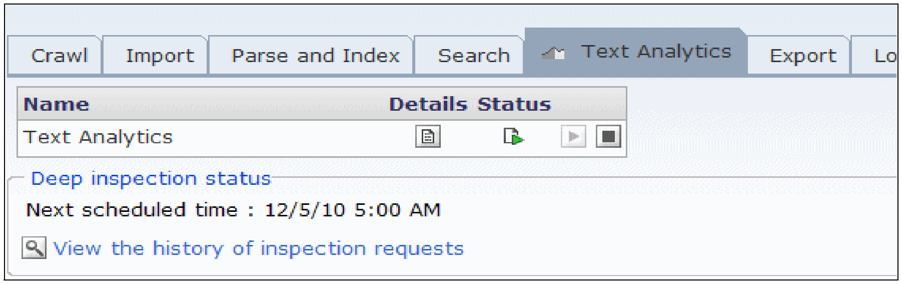
1. From the administration console, click **Collections** in the toolbar.
2. In the Collections view, locate the collection that you want to edit, and click the **Monitor** icon (Figure 10-56).



Crawl	Parse and Index	Search	Monitor	Edit
✓	✓	✓		

Figure 10-56 Collections View with editing and monitoring options

3. Click the **Text Analytics** tab (Figure 10-57), and click the **View the history of inspection requests** link to see all the requests made for deep inspections.



Name	Details	Status
Text Analytics		

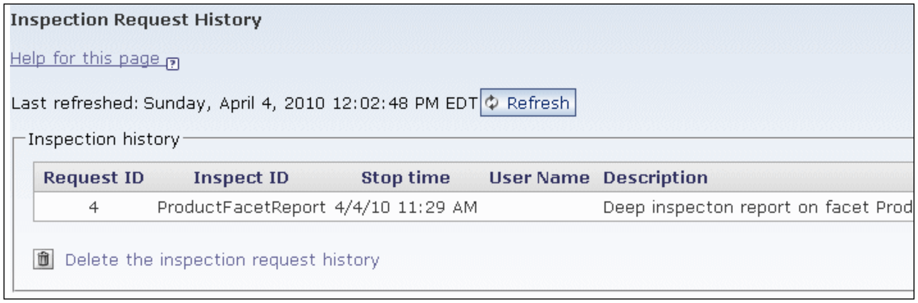
Deep inspection status

Next scheduled time : 12/5/10 5:00 AM

View the history of inspection requests

Figure 10-57 Viewing the deep inspection requests

Figure 10-58 shows all deep inspection requests that have been processed.



Inspection Request History

Help for this page

Last refreshed: Sunday, April 4, 2010 12:02:48 PM EDT

Inspection history

Request ID	Inspect ID	Stop time	User Name	Description
4	ProductFacetReport	4/4/10 11:29 AM		Deep inspecton report on facet Prod

Delete the inspection request history

Figure 10-58 Deep Inspection Request History window

10.7.7 Validating the deep inspection reports generation

You can check the directory for the deep inspection output to verify that the reports are being generated. The output contains two files in the XML format. The file name ending with `_log.xml` is a metadata file that contains information about the deep inspection request, such as the name of the report, and the time it is requested. Example 10-7 shows a metadata file.

Example 10-7 Metadata file about deep inspection request

```
<?xml version="1.0" encoding="UTF-8"?>
<Document Id="ProductFacetReport" Type="NORMAL">
  <Content Truncated="false"
Path="C:&#x5C;Export&#x5C;DeepInspectionReports&#x5C;20100404112938&#x5C;ProductFacet
Report.xml" Encoded="false"></Content>
  <Metadata>
    <Fields>
      <Field Name="QueryText">*:*</Field>
      <Field Name="Facet.1">Product</Field>
      <Field Name="Facet.1.Id">$.product</Field>
      <Field Name="Facet.1.TaxonomyType">keywords</Field>
      <Field Name="SortKey">frequency</Field>
      <Field Name="ViewName">Facets</Field>
      <Field Name="StartedDateTime">2010.04.04 11:29:38 EDT</Field>
      <Field Name="CompletionDateTime">2010.04.04 11:29:38 EDT</Field>
      <Field Name="ResultCode">0</Field>
      <Field Name="NumberOfRecords">10</Field>
    </Fields>
    <Facets></Facets>
  </Metadata>
</Document>
```

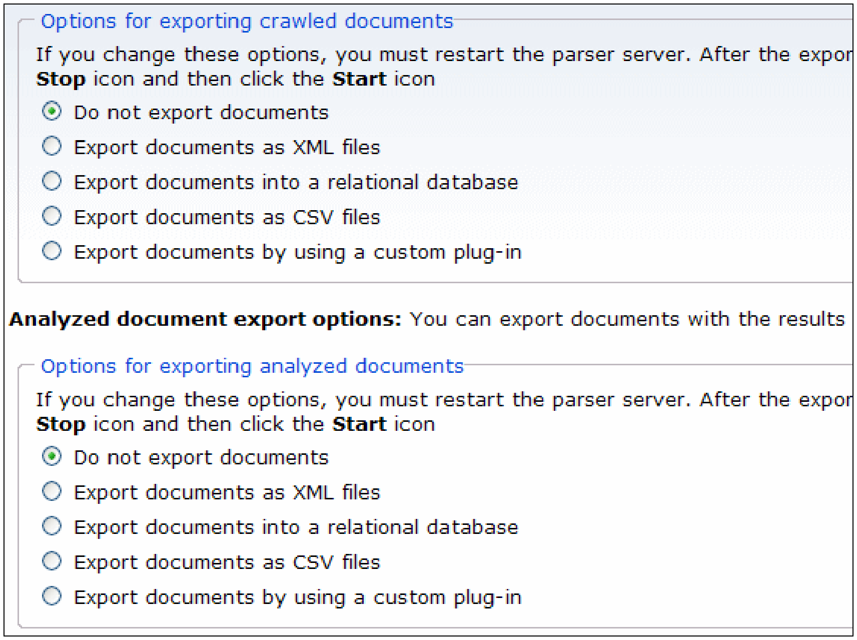
The second XML file is a deep inspection report that contains details about facets, counts, deviations, correlations, and other information. Example 10-8 shows a deep inspection report that contains the top 10 keywords for the Product facet. In this report, the Count attribute represents frequency. Because the deep inspection report is generated by using the Facets view, the Index attribute represents the correlation value. If the deep inspection report is generated from the Trends or Deviations views, the Index value represents the index value.

Example 10-8 A deep inspection report on the Product facet

```
<?xml version="1.0" encoding="UTF-8"?>
<Report Id="ProductFacetReport">
  <Record Rank="1" Count="101" Index="1.0">
    <Facet dimension="Facet.1">vanilla ice cream</Facet>
  </Record>
  <Record Rank="2" Count="86" Index="1.0">
    <Facet dimension="Facet.1">orange juice</Facet>
  </Record>
  <Record Rank="3" Count="85" Index="1.0">
    <Facet dimension="Facet.1">chocolate ice cream</Facet>
  </Record>
  <Record Rank="4" Count="58" Index="1.0">
    <Facet dimension="Facet.1">pastry</Facet>
  </Record>
  <Record Rank="5" Count="58" Index="1.0">
    <Facet dimension="Facet.1">mint jelly</Facet>
  </Record>
  <Record Rank="6" Count="53" Index="1.0">
    <Facet dimension="Facet.1">fruit jelly</Facet>
  </Record>
  <Record Rank="7" Count="50" Index="1.0">
    <Facet dimension="Facet.1">N/A</Facet>
  </Record>
  <Record Rank="8" Count="49" Index="1.0">
    <Facet dimension="Facet.1">apple juice</Facet>
  </Record>
  <Record Rank="9" Count="43" Index="1.0">
    <Facet dimension="Facet.1">pine juice</Facet>
  </Record>
  <Record Rank="10" Count="41" Index="1.0">
    <Facet dimension="Facet.1">chocolate</Facet>
  </Record>
</Report>
```

10.8 Creating and deploying a custom plug-in

You can customize both the export and deep inspection features by using the plug-in capability. A custom plug-in is a Java program that you write that is applicable to your business rules and processing of the exported data. Customizing can be done at each of three stages of export and for deep inspection. Figure 10-59 on page 447 shows an example of how you can select to configure a custom export plug-in for crawled and analyzed documents.



Options for exporting crawled documents

If you change these options, you must restart the parser server. After the expor
Stop icon and then click the **Start** icon

● Do not export documents
○ Export documents as XML files
○ Export documents into a relational database
○ Export documents as CSV files
○ Export documents by using a custom plug-in

Analyzed document export options: You can export documents with the results

Options for exporting analyzed documents

If you change these options, you must restart the parser server. After the expor
Stop icon and then click the **Start** icon

● Do not export documents
○ Export documents as XML files
○ Export documents into a relational database
○ Export documents as CSV files
○ Export documents by using a custom plug-in

Figure 10-59 Option to configure a custom export plug-in

You can use this option to export data to different relational databases or in a different format. For example, on social websites or forums, each page contains multiple entries from different users. When Content Analytics crawls the content, a single web page is treated as a single HTML document, and thus the page is exported as a single crawled document. If you want to export multiple entries (from different users) as multiple documents, you can create a custom export plug-in at the crawled document stage to separate multiple entries as individual documents and export them individually.

You can also use the custom plug-in option of the deep inspection feature to export data to a relational database instead of to the file system. Then you can use the Cognos 8 Business Intelligence tool to generate reports or to do additional analysis. By default, the deep inspection feature exports the results to the file system. Similarly, you can create a custom plug-in to analyze and deliver daily deep inspection reports by email.

For information about creating and deploying a custom export or deep inspection plug-in, go to the IBM Content Analytics Information Center at the following address, and search on *creating and deploying a plug-in for exporting documents or deep inspection results*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

11

Configuring annotators

Facets represent important information that is extracted from your content. You use facets to navigate through and analyze different dimensions of your data. IBM Content Analytics populates facets by processing your content through the document processing pipeline, which consists of multiple annotators. This chapter provides an overview of the document processing pipeline and the annotators that are available from Content Analytics, with an emphasis on the custom annotator.

This chapter includes the following sections:

- Document processing pipeline and the annotators
- Custom annotators
- Validation

Annotators: For details about major annotators, see the following chapters:

- For Dictionary Lookup and Pattern Matcher annotators, see Chapter 7, “Performing content analysis” on page 279.
- For the Classification Module annotator, see Chapter 9, “Content analysis with IBM Classification Module” on page 357.

11.1 Document processing pipeline and the annotators

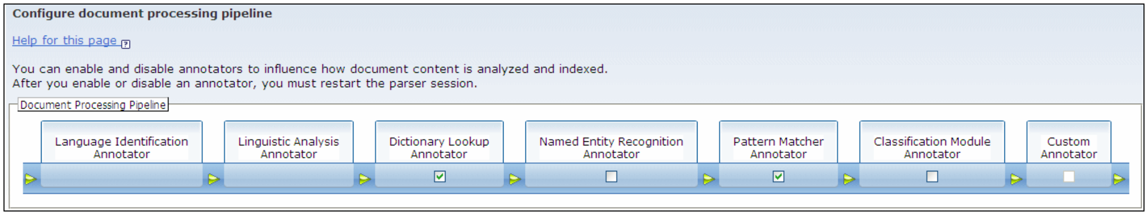
Content Analytics is a text analytics platform that conforms to the Unstructured Information Management Architecture (UIMA). Within Content Analytics is a UIMA-compliant document processing pipeline. As documents pass through the pipeline, they are examined by many annotators. Each annotator has a specific text analytic task to perform. This section provides an overview of the document processing pipeline and the various annotators within the pipeline.

11.1.1 UIMA document processing pipeline

Figure 11-1 shows a Content Analytics administration console of the UIMA pipeline that can be configured for each text analytics collection. The following annotators are executed in sequential order from left to right:

- Language Identification annotator
- Linguistic Analysis annotator
- Dictionary Lookup annotator
- Named Entity Recognition annotator
- Pattern Matcher annotator
- Classification Module annotator
- Custom annotators

The first two annotators are required to initiate document processing in the pipeline and thus cannot be disabled or removed. The last annotator, Custom annotator, is an optional annotator that you can develop and import into Content Analytics. The annotators in between the first two and the last annotator provide advanced text analytics functions that can be enabled and configured.



Configure document processing pipeline

Help for this page

You can enable and disable annotators to influence how document content is analyzed and indexed.
After you enable or disable an annotator, you must restart the parser session.

Document Processing Pipeline

Language Identification Annotator	Linguistic Analysis Annotator	Dictionary Lookup Annotator	Named Entity Recognition Annotator	Pattern Matcher Annotator	Classification Module Annotator	Custom Annotator
		☑	□	☑	□	□

Figure 11-1 Configure document processing pipeline window

11.1.2 Language Identification annotator

The Language Identification annotator identifies the language of each document. In many cases, documents collected from several sources can be in multiple languages. Before starting any text analytics, the language of a document must be determined. The Language Identification annotator is the first annotator to get control in the UIMA pipeline, and the annotator is used to identify the language of each document. The Language Identification annotator is always enabled automatically and cannot be disabled or removed, unless the collection is configured to support only one language.

When creating a collection, you can specify a list of candidate languages for the collection to avoid misdetection. If you want to create a collection and you know what the document languages are, for example English or German, you specify English and German in the list of candidate languages. After the Language Identification annotator runs, Content Analytics correctly identifies the language used for each document.

When creating a crawler for a collection, you can specify a particular language. For more information about creating a collection, see 4.3.2, "Creating a text analytics collection" on page 90. For more information about creating a crawler, see 4.3.3, "Defining and configuring a crawler" on page 123.

11.1.3 Linguistic Analysis annotator

The Linguistic Analysis annotator applies linguistic analysis for each document according to the corresponding language of the document. Based on the language identified from the Language Identification annotator, the Linguistic Analysis annotator segments a document into sentences and words. It finds the normalized form of these words and disambiguates the part of speech of each word.

For example, consider the following (short) document:

```
He is a software engineer.
```

After running the Linguistic Analysis annotator, Content Analytics knows that the document has one sentence and five words, and the second word `is` is a verb with a normalized form of `be`. The linguistic normal form of a word is called its *lemma*. The process in its entirety is referred to as *lexical analysis*. The Linguistic Analysis annotator is always enabled automatically and cannot be disabled or removed.

The Linguistic Analysis annotator contains dictionaries for all of the supported languages. If you want to add more specific words, for example product names, you can add them by supplying your own custom dictionary using the Dictionary Lookup annotator. The dictionaries for the Dictionary Lookup annotator are automatically applied to the Linguistic Analysis annotator. For more information about additional dictionaries, see 7.3, "Configuring the Dictionary Lookup annotator" on page 299.

11.1.4 Named Entity Recognition annotator

The Named Entity Recognition annotator identifies terms that signify persons, locations, and organizations in a document. For example, assume that the Named Entity Recognition annotator examines the sentence “`Gregory works at IBM in California`.” The annotator captures the name of the person `Gregory`, the name of organization `IBM`, and the name of location `California`. The named entities are displayed in the predefined Named Entity facet.

The Named Entity Recognition annotator supports the following languages:

- English
- French
- German
- Japanese
- Spanish

The Named Entity Recognition annotator is disabled by default. You can enable the annotator by using the administration console. You need to rebuild the index in order for the new entities to be extracted.

> **Important:** You cannot customize or configure the Named Entity Recognition annotator.

11.1.5 Dictionary Lookup and Pattern Matcher annotators

The Dictionary Lookup annotator and Pattern Matcher annotator are two key annotators that you can use to perform content analysis and to gain insight from your content. These annotators are fundamental tools to fine-tune and improve your content analysis. For detailed information about these annotators, see Chapter 7, “Performing content analysis” on page 279.

You can find examples of their usage in the following sections:

- 7.3, “Configuring the Dictionary Lookup annotator” on page 299
- 7.4, “Configuring the Pattern Matcher annotator” on page 309

11.1.6 Classification Module annotator

The Classification Module annotator is an important annotator that you can use to classify your documents, to get the necessary metadata, and to help you improve your content analysis. For detailed information about this annotator and its usage, see Chapter 9, “Content analysis with IBM Classification Module” on page 357.

11.2 Custom annotators

Content Analytics is packaged with useful annotators for advanced text analytics. If you want to add your own, you can develop a custom annotator and add it to the Content Analytics document processing pipeline.

The Content Analytics document processor is based on the Apache UIMA framework and supports the integration of custom UIMA-compliant annotators. Content Analytics supports the import of a custom Text Analysis Engine (TAE) as a processing engine archive (PEAR) file. A single TAE can have one or more annotators working together to achieve a specific text analytic objective. For example, an objective might be to identify and extract the names of people who have collaborated on a specific business deal.

A PEAR file is a standard deployment package format for UIMA components. It is the vehicle that packages a TAE and makes it available for import into Content Analytics. The development of individual annotators, their assembly into a TAE, and subsequent deployment as a PEAR file can be performed by using the Apache UIMA software development kit (SDK). The Apache UIMA SDK is available at the following web address:

http://uima.apache.org/downloads.cgi

In addition, Content Analytics provides the LanguageWare Resource Workbench, which is a flexible and easy-to-use development environment for the creation of sophisticated custom annotators in various languages. The deployment vehicle of the LanguageWare Resource Workbench is the PEAR file.

Single versus multiple PEAR files: Content Analytics only supports the association of a single custom PEAR file with a text analytics collection. To use multiple PEAR files with a single collection, you must reassemble them into a single PEAR file by using the LanguageWare Resource Workbench or other tools.

11.2.1 LanguageWare Resource Workbench

IBM LanguageWare is a technology that provides a full range of text analysis functionality. LanguageWare is used extensively throughout IBM products and services. Content Analytics uses LanguageWare technologies to process and understand natural language text.

LanguageWare Resource Workbench is a powerful tool for creating your own custom text analytics annotators. With this tool, users can easily create annotators for the Content Analytics applications without having to write any code. The custom annotator created in LanguageWare Resource Workbench can then be exported to be used with Content Analytics.

LanguageWare Resource Workbench is an Eclipse-based, integrated development environment for creating dictionaries, rules, and UIMA annotators. With LanguageWare Resource Workbench, you can perform the following tasks:

- Build language and domain resources into dictionaries.
- Develop rules to spot facets, entities, and relationships by using a simple drag-and-drop paradigm.
- Create UIMA annotators for annotating text with dictionaries and rules.
- Annotate text and browse the contents of each annotation.
- Export a UIMA text analysis engine as a UIMA PEAR file so that the same annotator can be run outside of the Workbench.
- Export a UIMA text analysis engine directly to Content Analytics. The exported engine can be automatically added as the custom annotator stage of the document processing pipeline in a collection.

The LanguageWare Resource Workbench is available from IBM alphaWorks® under a 90-day early release license. If you purchase Content Analytics, you are entitled to one LanguageWare license available at no charge and associated use of the LanguageWare Resource Workbench. You can download and try the LanguageWare Resource Workbench. For details, see “Text Analytics Tools and Runtime for IBM LanguageWare” in IBM alphaWorks at the following address:

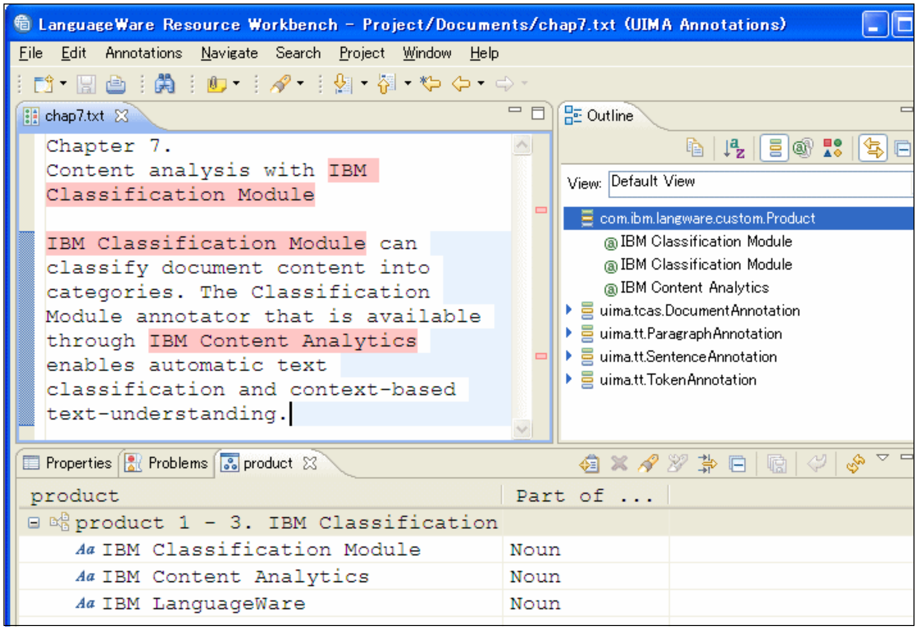
http://www.alphaworks.ibm.com/tech/lrw/

Custom annotator use cases

You can build the following types of custom annotators by creating LanguageWare resources (LanguageWare dictionaries and rules):

- Dictionary-based annotation

  You can create dictionaries that contain word forms and associated information. For example, when you have a list of product names, you can easily create a product name dictionary that contains the product name and associated information such as the product ID and product alias name. The LanguageWare Resource Workbench supports the annotation of documents based on the words that match the words in dictionaries. The document terms that match are highlighted by the editor (Figure 11-2). You can test your dictionaries in real time by annotating text documents and browsing the contents of each annotation.


LanguageWare Resource Workbench - Project/Documents/chap7.txt (UIMA Annotations)
File Edit Annotations Navigate Search Project Window Help
chap7.txt
Outline
View: Default View
com.ibm.langware.custom.Product
IBM Classification Module
IBM Classification Module
IBM Content Analytics
uima.tcas.DocumentAnnotation
uima.tt.ParagraphAnnotation
uima.tt.SentenceAnnotation
uima.tt.TokenAnnotation
Properties
Problems
product

```
Chapter 7.
Content analysis with IBM
Classification Module

IBM Classification Module can
classify document content into
categories. The Classification
Module annotator that is available
through IBM Content Analytics
enables automatic text
classification and context-based
text-understanding.
```

product	Part of ...
product 1 - 3. IBM Classification	
Aa IBM Classification Module	Noun
Aa IBM Content Analytics	Noun
Aa IBM LanguageWare	Noun

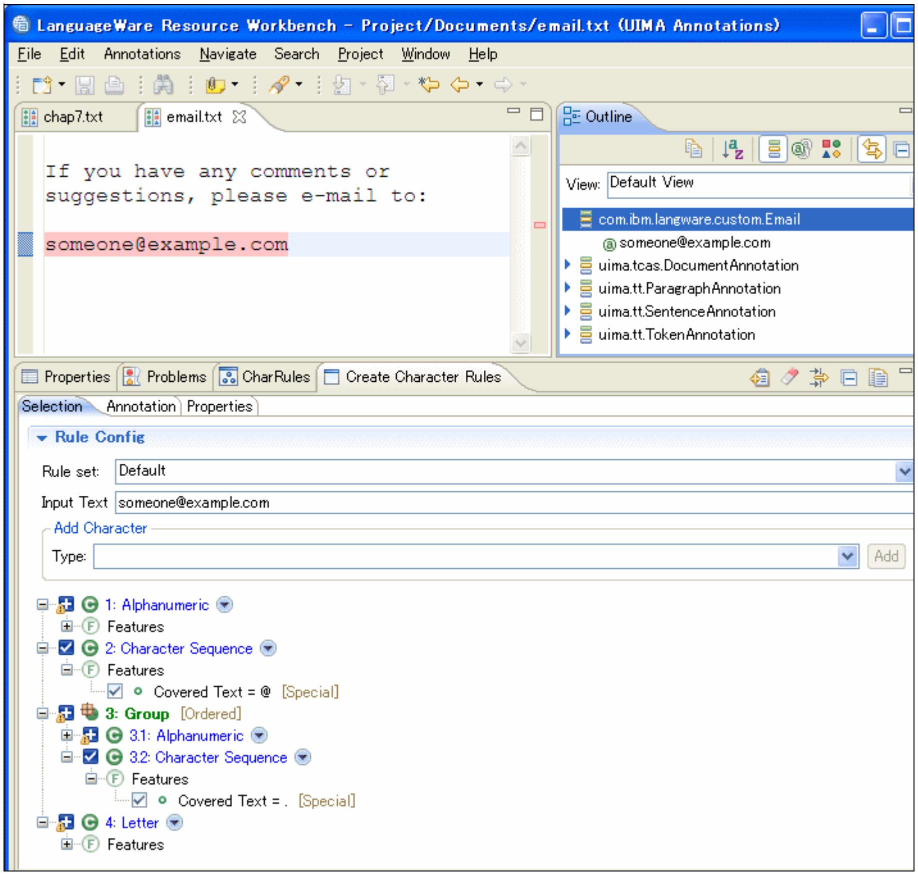
Figure 11-2 Annotating product names by using a custom dictionary

- Character rule-based annotation

  You can define patterns of characters that represent specific types of information. Character rules are usually written by dragging a character sequence on the Character Rules Editor. Character rules can be also defined manually with the regular expression syntax. It is useful to capture the following information for example:

  – Telephone number
  – Email address
  – Uniform Resource Locator (URL)
  – International Standard Book Number (ISBN)

  Figure 11-3 shows the workbench editing the rules for an email address and the annotated document within the editor.


LanguageWare Resource Workbench - Project/Documents/email.txt (UIMA Annotations)
File Edit Annotations Navigate Search Project Window Help
chap7.txt
email.txt
Outline
Properties
Problems
CharRules
Create Character Rules
Selection
Annotation
Properties
Rule Config
Add Character
Add

```
If you have any comments or
suggestions, please e-mail to:

someone@example.com
```

View: Default View

- com.ibm.langware.custom.Email
  - someone@example.com
- uima.tcas.DocumentAnnotation
- uima.tt.ParagraphAnnotation
- uima.tt.SentenceAnnotation
- uima.tt.TokenAnnotation

Rule set: Default

Input Text: someone@example.com

Type:

```
1: Alphanumeric
   Features
2: Character Sequence
   Features
      Covered Text = @  [Special]
3: Group  [Ordered]
   3.1: Alphanumeric
   3.2: Character Sequence
      Features
         Covered Text = .  [Special]
4: Letter
   Features
```

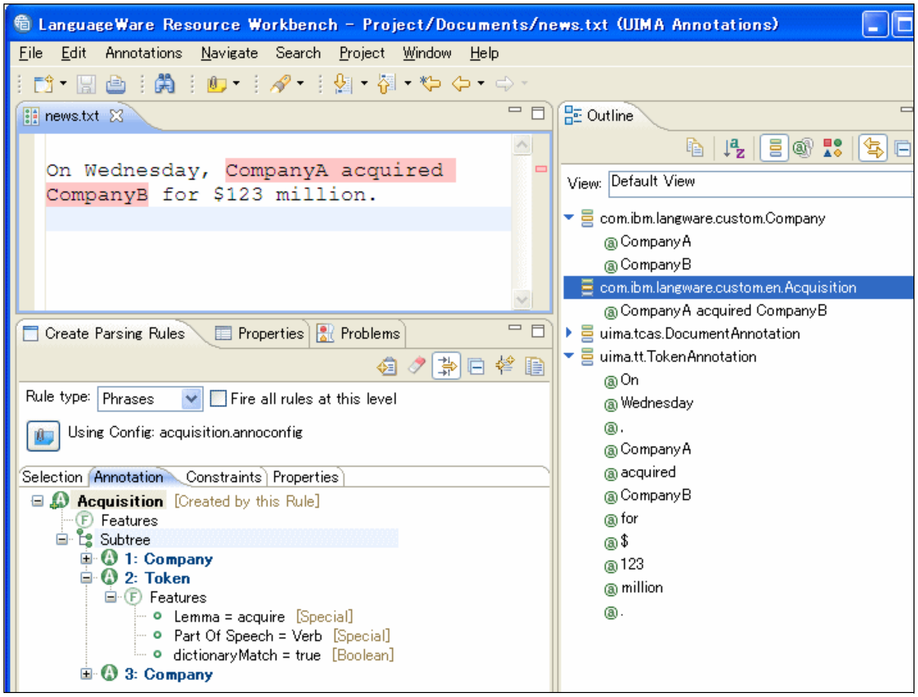
Figure 11-3 Annotating an email address using character rules

- ► Rule annotation

  You can further define rules to analyze tokens and other UIMA annotations created by previous annotators (such as dictionary-based and character rule-based annotators). You can also define them to create new annotations based on identifying patterns in those annotations. Rules are written by dragging a section of text to the Rules Editor, for example.

  Figure 11-4 shows an example of rule annotation. In this example, `CompanyA` and `CompanyB` are dictionary-based annotations created by the previous annotator. The rule represents the following sequence of annotations:

  – A `Company` annotation from a dictionary of company names
  – A `Token` annotation, which has a lemma of `acquire`
  – A `Company` annotation from a dictionary of company names



On Wednesday, CompanyA acquired CompanyB for $123 million.

Rule type: Phrases
□ Fire all rules at this level

Using Config: acquisition.annoconfig

```
Acquisition [Created by this Rule]
  Features
  Subtree
    1: Company
    2: Token
      Features
        Lemma = acquire [Special]
        Part Of Speech = Verb [Special]
        dictionaryMatch = true [Boolean]
    3: Company
```

View: Default View

```
com.ibm.langware.custom.Company
  CompanyA
  CompanyB
com.ibm.langware.custom.en.Acquisition
  CompanyA acquired CompanyB
uima.tcas.DocumentAnnotation
uima.tt.TokenAnnotation
  On
  Wednesday
  ,
  CompanyA
  acquired
  CompanyB
  for
  $
  123
  million
  .
```

Figure 11-4 Annotating an acquisition phrase using rules

By dragging the text “`CompanyA acquired CompanyB`” to the Rules Editor, you can label it as `Acquisition annotation`. Rules can be made either more generic or more specific by tuning the match criteria of the annotations.

You can also create aggregate rules that depend on previous rules. For example, RuleA identifies `Acquisition annotation` (shown in the previous example), and RuleB identifies `Acquisition Cost annotation`. They are represented as `[Acquisition annotation] + [any token] + [Currency annotation]` in the same sentence, which means that RuleB depends on RuleA.

- Custom annotation

  Although the LanguageWare Resource Workbench provides a wide range of capabilities for analysis, it is sometimes useful to use an annotator written for a special purpose. You can use any UIMA-compliant annotator within the LanguageWare Resource Workbench.

You can easily export annotators as a UIMA PEAR file, which can be run in any UIMA-compliant application. The exported PEAR file contains all required resources, libraries, and configurations to run the annotators outside the LanguageWare Resource Workbench environment. The LanguageWare Resource Workbench also supports the integration of an exported PEAR file into the Content Analytics document processing pipeline.

For more information, see the help contents in the LanguageWare Resource Workbench and the training materials that are available from IBM alphaWorks:

http://www.alphaworks.ibm.com/

Creating a connection to Content Analytics server

After you create your UIMA annotators in the LanguageWare Resource Workbench, you can export them as a PEAR for use within Content Analytics. To install the PEAR directly to Content Analytics, create a connection to Content Analytics server by using the following steps:

1. Select **File** → **New** → **IBM Content Analytics Server Connection**.
2. In the Content Analytics Server connection window, specify the name of the connection file and the folder in which you want to create the connection file. Click **Finish**.

3. In the Content Analytics Server Connection editor (Figure 11-5), where the connection file is displayed, specify the Content Analytics server information:

   a. Enter the fully qualified host name of the Content Analytics server.

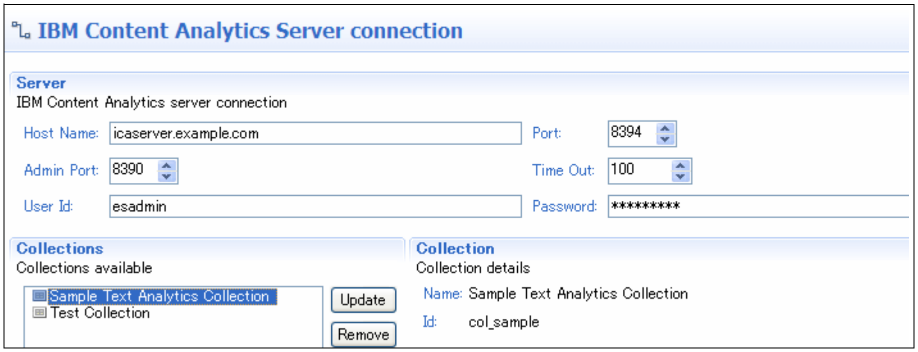
   b. Specify the search server port number as `Port`. By default, the port number is 8394.

   c. Specify the administration console port number as `Admin Port`. The default port number is 8390.

   d. Enter the administrator user name.

   e. Enter the password of the administrator user.

   f. Click **Update**. The Workbench connects to the Content Analytics server and shows a list of collections that are available on the server.



IBM Content Analytics Server connection

Server

IBM Content Analytics server connection

Host Name: icaserver.example.com

Port: 8394

Admin Port: 8390

Time Out: 100

User Id: esadmin

Password: *********

Collections

Collections available

- Sample Text Analytics Collection
- Test Collection

Collection

Collection details

Name: Sample Text Analytics Collection

Id: col_sample

Figure 11-5 Content Analytics Server Connection editor

4. Save the connection file.

Installing a custom annotator to Content Analytics

After configuring the connection file, you can install a custom annotator to Content Analytics and use it with a text analytics collection. Using the custom annotator to capture information about the acquisition of the company (Figure 11-4 on page 457), export the custom annotator into Content Analytics and configure the text processing options:

1. Select **File** → **Export**.
2. Select **IBM LanguageWare** → **LanguageWare UIMA Pipeline to IBM Content Analytics** option. Click **Next**.

3. In the Export PEAR window (Figure 11-6), configure the PEAR package:

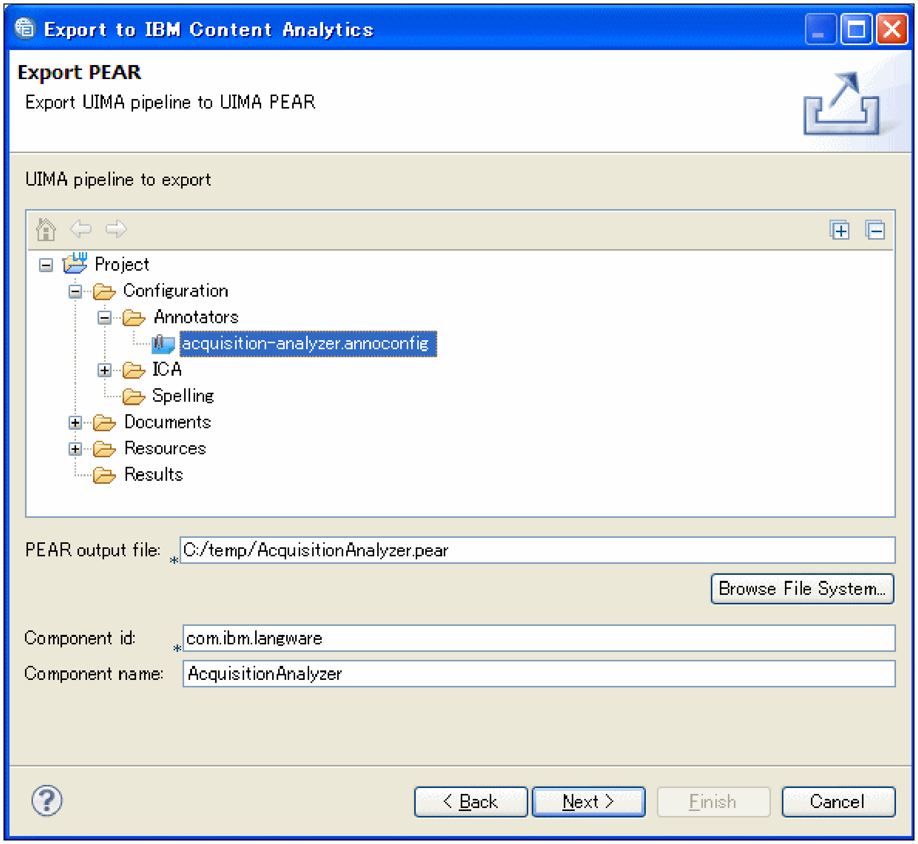
   a. Browse your workspace and select the LanguageWare UIMA Pipeline configuration file (`.annoconfig` file) that you want to export.

   b. For PEAR output file, enter the PEAR file path that will be created on the local machine. The PEAR file has the `.pear` extension.

   c. For Component id, enter an ID that is a unique identifier of the component. The ID must follow the Java package naming convention.

   d. For Component name, enter a name using the Java class naming convention.

   e. Click **Next**.



Export to IBM Content Analytics

Export PEAR

Export UIMA pipeline to UIMA PEAR

UIMA pipeline to export

- Project
  - Configuration
    - Annotators
      - acquisition-analyzer.annoconfig
    - ICA
    - Spelling
  - Documents
  - Resources
  - Results

PEAR output file: * C:/temp/AcquisitionAnalyzer.pear

Component id: * com.ibm.langware

Component name: AcquisitionAnalyzer

Figure 11-6 Specifying a PEAR package configuration

4. In the File Selection window (Figure 11-7), browse your workspace and select the Content Analytics Server connection file (`.icaconfig` file) that defines the Content Analytics server information. Click **Next**.

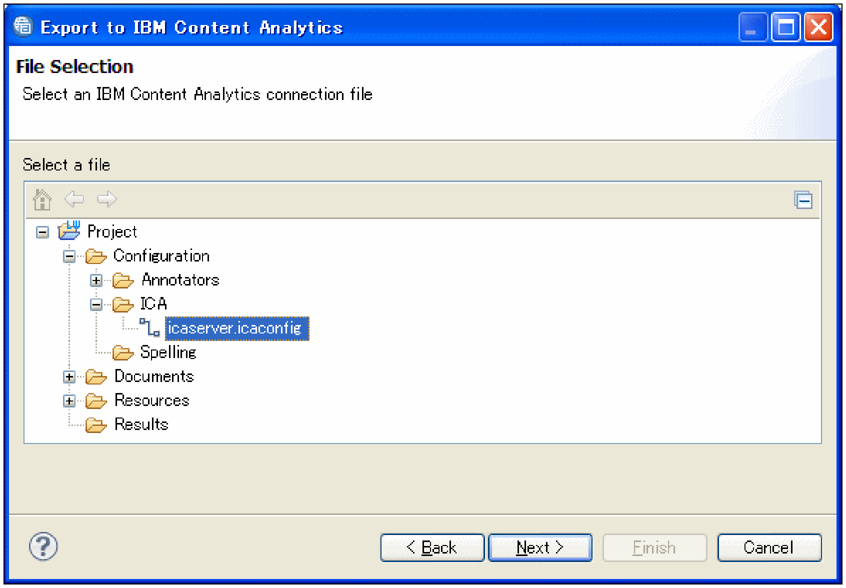


Figure 11-7   Selecting a connection file

5. In the Collections window (Figure 11-8), select a text analytics collection that you want to associate the text analysis engine with. Click **Next**.

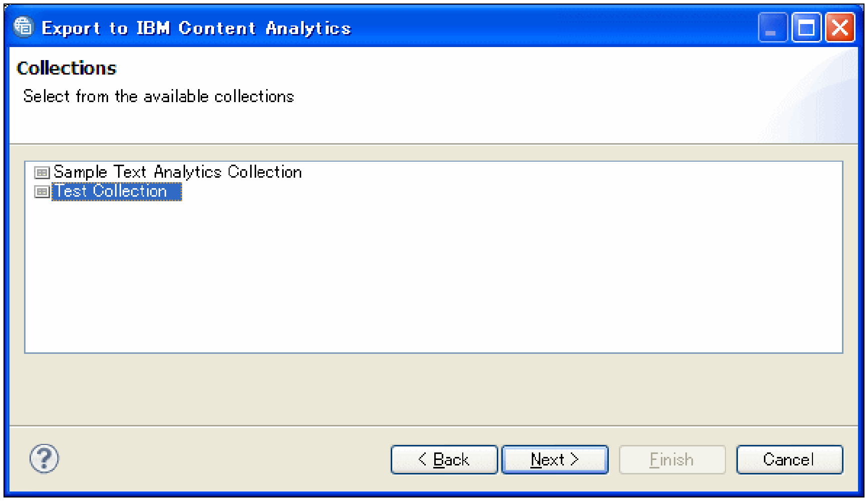
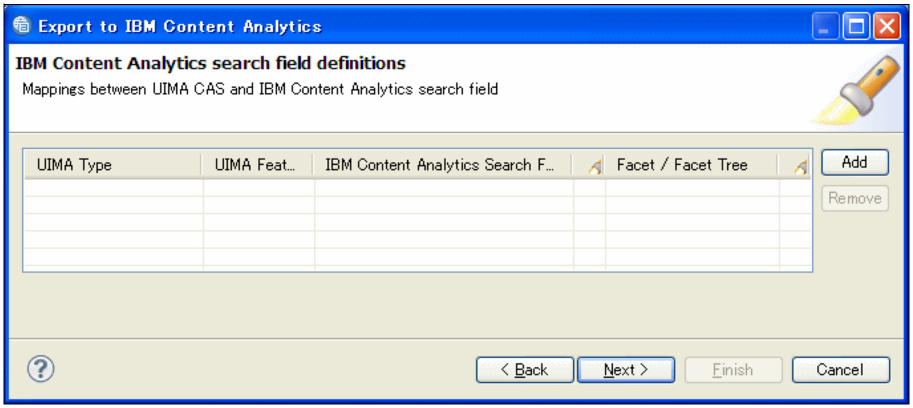

Export to IBM Content Analytics
Collections
Select from the available collections
Sample Text Analytics Collection
Test Collection
< Back
Next >
Finish
Cancel

Figure 11-8   Selecting a text analytics collection

6. In the Content Analytics search field definitions window (Figure 11-9), specify which UIMA Annotations are indexed by Content Analytics and available as search fields and facets. Click **Add** to create a mapping between a UIMA annotation and a search field.



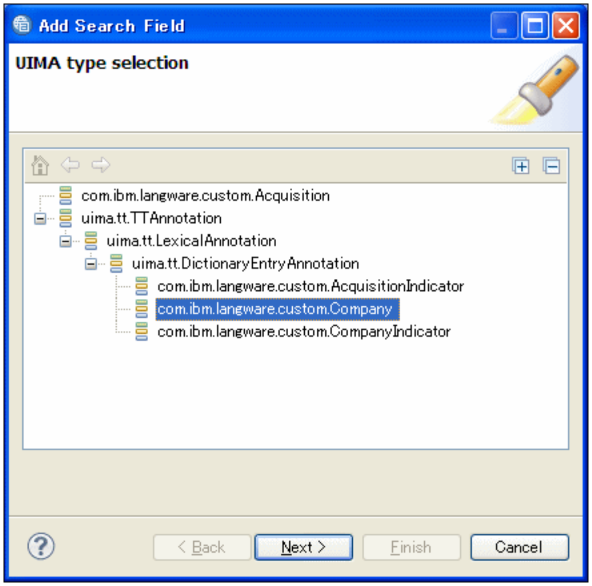
IBM Content Analytics search field definitions

Mappings between UIMA CAS and IBM Content Analytics search field

UIMA Type	UIMA Feat...	IBM Content Analytics Search F...	Facet / Facet Tree

Figure 11-9 Content Analytics search field definitions window

7. In the UIMA type selection window that shows the UIMA types tree (Figure 11-10), select the UIMA type that is associated with the Content Analytics search field. In this example, **com.ibm.langware.custom. Company** a dictionary-based annotation type is selected. Click **Next**.


Add Search Field
com.ibm.langware.custom.Acquisition
uima.tt.TTAnnotation
uima.tt.LexicalAnnotation
uima.tt.DictionaryEntryAnnotation
com.ibm.langware.custom.AcquisitionIndicator
com.ibm.langware.custom.Company
com.ibm.langware.custom.CompanyIndicator
< Back
Next >
Finish
Cancel

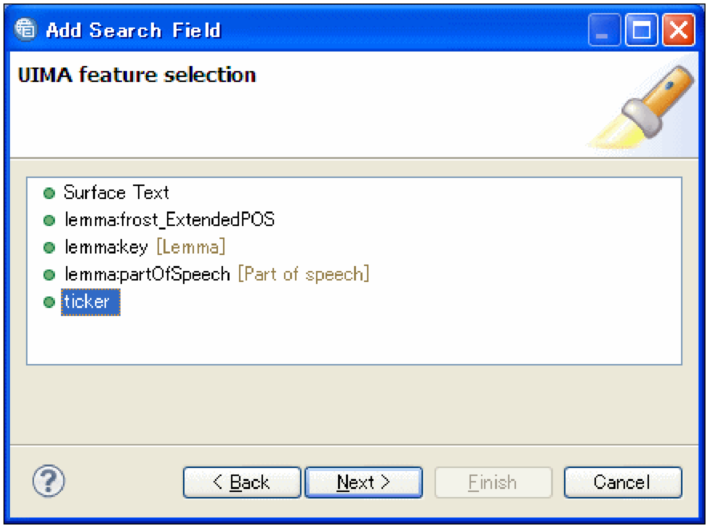
UIMA type selection

Figure 11-10 Selecting a UIMA type

The window now shows a list of UIMA features of the selected UIMA type. The following features are important:

- `Surface Text` is the annotated text itself (the annotation span of text).
- `lemma:key` is the lemma (normalized form) of the annotated word. It is usually used for dictionary-based annotation.

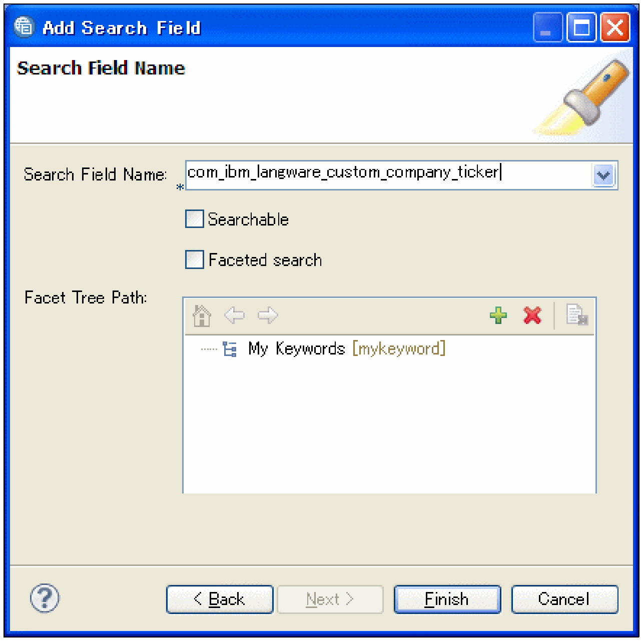
8. In the UIMA feature selection window (Figure 11-11), select the UIMA feature. In this figure, the **ticker** feature is selected. Then click **Next**.



- Surface Text
- lemma:frost_ExtendedPOS
- lemma:key [Lemma]
- lemma:partOfSpeech [Part of speech]
- ticker

Figure 11-11 Selecting a UIMA feature

9. In the Search Field Name window (Figure 11-12), specify a search field name that is associated with the UIMA type and feature. By default, it uses the type name followed by the feature name. You can change the name if necessary. You can also select an existing search field that is available in the text analytics collection.



Search Field Name

Search Field Name: * com_ibm_langware_custom_company_ticker

□ Searchable

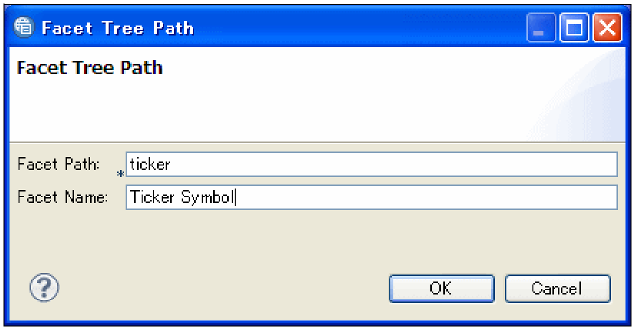
□ Faceted search

Facet Tree Path:

My Keywords [mykeyword]

Figure 11-12 Entering the search field name

If you also want to associate with a facet, select the **Faceted search** option and specify a facet. You can create a facet or select an existing facet that is defined in the text analytics collection. Figure 11-13 shows the window to create a Ticker Symbol facet.



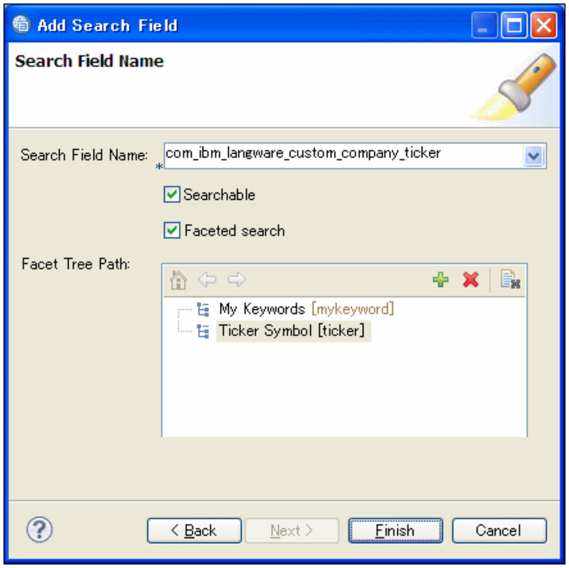
Facet Tree Path

Facet Path: * ticker

Facet Name: Ticker Symbol

Figure 11-13 Creating a facet

After specifying the search field and the facet (Figure 11-14), click **Finish**.



Add Search Field

Search Field Name

Search Field Name: * com_ibm_langware_custom_company_ticker

☑ Searchable

☑ Faceted search

Facet Tree Path:

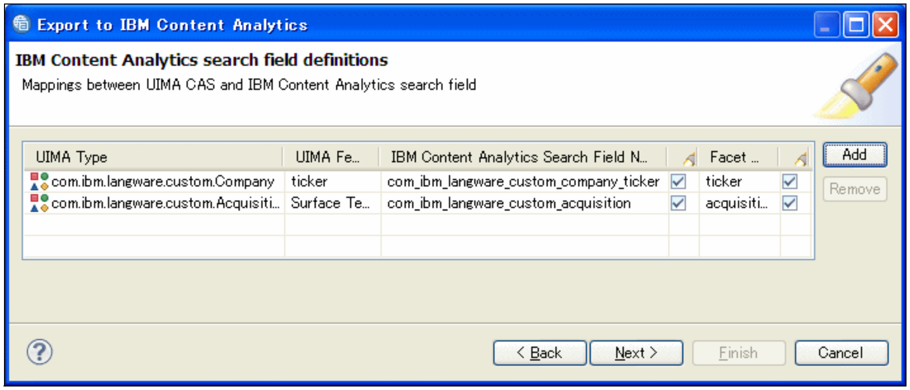
- My Keywords [mykeyword]
- Ticker Symbol [ticker]

Figure 11-14 After configuring the search field and the facet

10. Repeat step 6 on page 462 through step 9 on page 464 until you add all required mappings.

    Figure 11-15 shows the list of added definitions. In this example, Two UIMA annotations are added. One is `com.ibm.langware.custom.Company:ticker` that is associated with the Ticker Symbol facet. The other is `com.ibm.langware.custom.Acquisition` that is associated with the Acquisition Phrase facet.

    Click **Next**.



Export to IBM Content Analytics

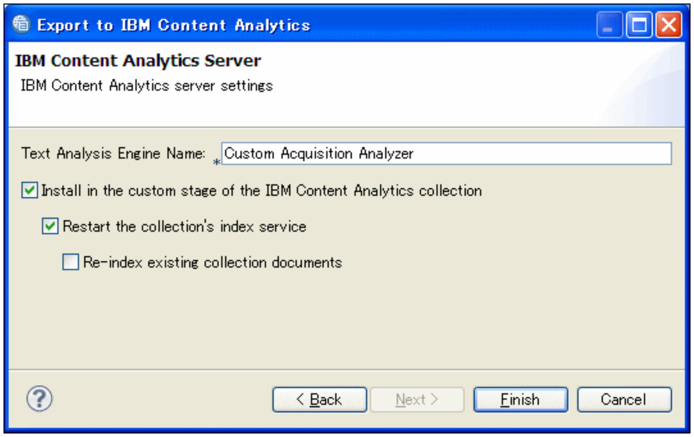
IBM Content Analytics search field definitions

Mappings between UIMA CAS and IBM Content Analytics search field

UIMA Type	UIMA Fe...	IBM Content Analytics Search Field N...		Facet ...	
com.ibm.langware.custom.Company	ticker	com_ibm_langware_custom_company_ticker	☑	ticker	☑
com.ibm.langware.custom.Acquisiti...	Surface Te...	com_ibm_langware_custom_acquisition	☑	acquisiti...	☑

Figure 11-15 Content Analytics search field definitions showing the added definitions

11. In the Content Analytics Server window (Figure 11-16), enter the **Text Analysis Engine Name**. This option specifies the name that the Content Analytics server assigns to the text analysis engine. This name must be unique on the server. Then click **Finish**.



IBM Content Analytics Server

IBM Content Analytics server settings

Text Analysis Engine Name: * Custom Acquisition Analyzer

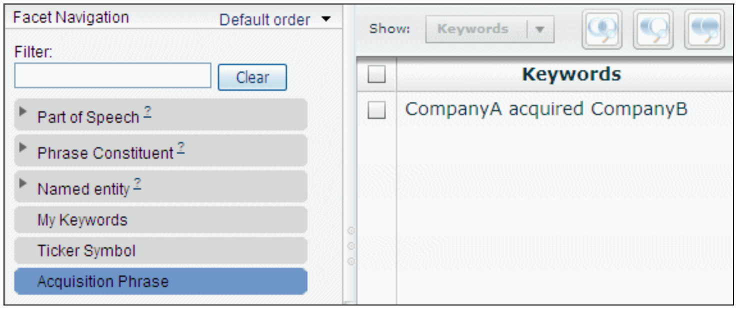
☑ Install in the custom stage of the IBM Content Analytics collection

☑ Restart the collection's index service

☐ Re-index existing collection documents

Figure 11-16 Specifying the Text Analysis Engine Name

After you complete these steps, the LanguageWare Resource Workbench exports the PEAR file and installs it to the Content Analytics server. This process might take some time. After crawling sample documents, you can see the Acquisition Phrase facet in the Facet view as shown in Figure 11-17.



☐	Keywords
☐	CompanyA acquired CompanyB

Figure 11-17 Facet View showing the Acquisition Phrase facet from the installed text analysis engine

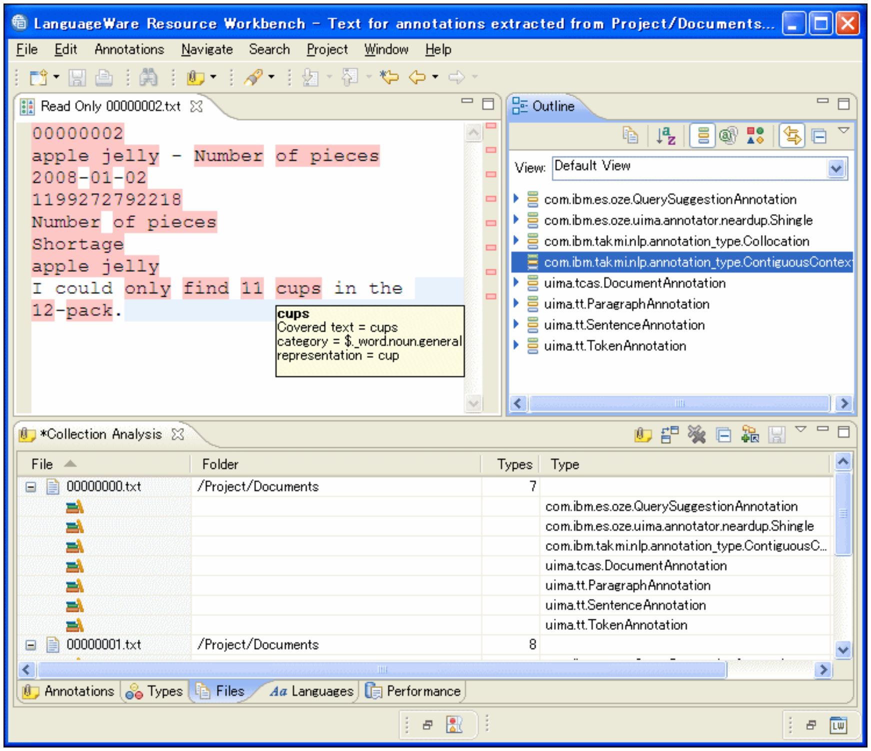
Analyzing documents using a Content Analytics pipeline

The LanguageWare Resource Workbench can pass documents in the Workbench project to a Content Analytics server to be annotated by the document processing pipeline associated with a collection. The LanguageWare Resource Workbench shows the resulting UIMA Annotations that are returned from Content Analytics. You can see UIMA annotations that come from a Content Analytics document processing pipeline the same as when you run UIMA annotators on the local environment.

To annotate documents on a Content Analytics pipeline, follow these steps:

1. Select a text file or a folder that contains text files.
2. Right-click and select **Analyze Collection with IBM Content Analytics**.
3. Browse your workspace and select the Content Analytics Server Connection configuration file (`.icaconfig` file). Click **Next**. If you have not created a connection file, follow the steps in “Creating a connection to Content Analytics server” on page 458.
4. Select a text analytics collection that contains a document processing pipeline that you want to run. Click **Finish**.

Figure 11-18 shows annotation results that are returned by the document processing pipeline in the Sample Text Analytics Collection. (See Chapter 7, “Performing content analysis” on page 279.) You can save the annotation results and compare them later.


LanguageWare Resource Workbench - Text for annotations extracted from Project/Documents...
File Edit Annotations Navigate Search Project Window Help
Read Only 00000002.txt
Outline
View: Default View
*Collection Analysis
Annotations Types Files Languages Performance

```
00000002
apple jelly - Number of pieces
2008-01-02
1199272792218
Number of pieces
Shortage
apple jelly
I could only find 11 cups in the
12-pack.
```

cups
Covered text = cups
category = $._word.noun.general
representation = cup

- com.ibm.es.oze.QuerySuggestionAnnotation
- com.ibm.es.oze.uima.annotator.neardup.Shingle
- com.ibm.takmi.nlp.annotation_type.Collocation
- com.ibm.takmi.nlp.annotation_type.ContiguousContex
- uima.tcas.DocumentAnnotation
- uima.tt.ParagraphAnnotation
- uima.tt.SentenceAnnotation
- uima.tt.TokenAnnotation

File	Folder	Types	Type
00000000.txt	/Project/Documents	7	
			com.ibm.es.oze.QuerySuggestionAnnotation
			com.ibm.es.oze.uima.annotator.neardup.Shingle
			com.ibm.takmi.nlp.annotation_type.ContiguousC...
			uima.tcas.DocumentAnnotation
			uima.tt.ParagraphAnnotation
			uima.tt.SentenceAnnotation
			uima.tt.TokenAnnotation
00000001.txt	/Project/Documents	8	

Figure 11-18 Showing annotation results that are returned by Content Analytics pipeline

Switching the CAS view: If you install custom annotators to a text analytics collection using the LanguageWare Resource Workbench, the resulting Common Analysis Structure (CAS) contains two CAS views: default view and lrw-view. The *default view* contains annotations derived from Content Analytics built-in annotators. The *lrw-view* has annotations coming from the custom annotators. You can switch to any of the CAS views by using the selection box in the Outline view.

Maintenance and validation

One of the most preferred practices in custom annotator development is to develop the annotator iteratively. Start with a small document set and create a small set of resources first. Then check the results to ensure that they are as expected. Continue to enhance the annotators by extending the resources and adding rules. Repeat these steps, checking your results as you go.

The LanguageWare Resource Workbench supports the iterative process. This process includes annotating a collection of documents, saving annotation results, comparing annotation results, and checking performance numbers.

Another preferred practice is to keep the rules simple and clean. Do not try to stretch a rule to capture everything you want. Rules that are too complex are difficult to maintain. Such rules might cause a performance degradation.

After deploying a PEAR file, verify that the result is what you expected. You can see the annotation results that are returned by the Content Analytics pipeline in the LanguageWare Resource Workbench. This result is useful for low-level investigation. You also can use Real-time natural language processing (NLP) to validate the results. With Real-time NLP, you can see facet values interactively with your documents that were used with the LanguageWare Resource Workbench when creating your annotators. For more information about Real-time NLP, see 11.3.1, “Real-time NLP” on page 476.

11.2.2 Creating custom annotators using the Apache UIMA SDK

UIMA is an open standard and open source. You can use the Apache UIMA SDK to implement a custom annotator in Java if you want your own analytics logic and resources.

Creating a custom annotator

To create a custom annotator, follow these steps:

1. Install the Apache UIMA SDK. The Apache UIMA SDK is available on the Apache UIMA site at the following address:

   http://uima.apache.org/downloads/

2. Implement the analysis logic (annotator) by using the UIMA SDK.
3. Package the analysis engine as a PEAR file, which is a standard package format for UIMA components.

Content Analytics contains a Regular Expression annotator that is a part of the UIMA available annotators. The sample files are in the `ES_INSTALL_ROOT/packages/uima/apache-uima-regex` directory:

RegExAnnotator.pear
Contains all necessary libraries, resources, and configurations.

sample_regex_cas2index.xml
Contains mappings between CAS and the Content Analytics index.

sample_regex_field.xml
Contains search field definitions for a collection.

For more information about the sample files, go to the IBM Content Analytics Information Center at the following address, and search on *using the sample UIMA Regular Expression Annotator*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Importing a PEAR file into Content Analytics

After you create the PEAR file, you can import it into Content Analytics and use it with a text analytics collection. You typically perform the following steps:

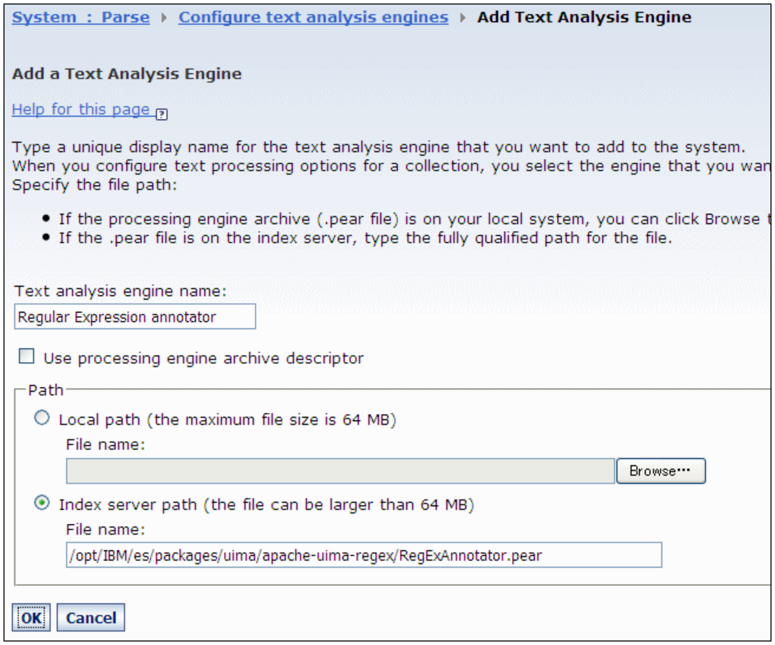
1. Upload the PEAR file to the Content Analytics system.
2. Configure the text processing options:
   a. Associate your engine with a collection. In many cases, you do not need to specify the CAS view name when selecting a text analysis engine.
   b. Create a CAS to Content Analytics index mapping file, and upload it. For details about mapping, go to the IBM Content Analytics Information Center at the following address, and search on *creating the common analysis structure to index mapping file*:

      http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

c. Define the search fields for the collection and associate the search fields to the facet path.

Using the Regular Expression annotator, use the custom annotator with Content Analytics as explained in the following steps:

1. From the administration console, select the **System** tab.
2. Click **Edit**.
3. Select **Parse** → **Configure text analysis engines**.
4. Select **Add Text Analysis Engine**.
5. Specify the configuration settings (Figure 11-21):
   a. Enter the text analysis engine name, which must be a unique name.
   b. Select the **Use processing engine archive descriptor** option. If this option is selected, the PEAR descriptor is used to set up an internal resource manager and localized class path for the custom text analysis engine. It is used to avoid conflicts of resources. You must select this option when using the PEAR file that is exported from the LanguageWare Resource Workbench.
   c. Specify the PEAR file that you want to add.
   d. Click **OK**.


System : Parse ▸ Configure text analysis engines ▸ Add Text Analysis Engine

Add a Text Analysis Engine

Help for this page

Type a unique display name for the text analysis engine that you want to add to the system.
When you configure text processing options for a collection, you select the engine that you wan
Specify the file path:

- If the processing engine archive (.pear file) is on your local system, you can click Browse t
- If the .pear file is on the index server, type the fully qualified path for the file.

Text analysis engine name:
Regular Expression annotator

□ Use processing engine archive descriptor

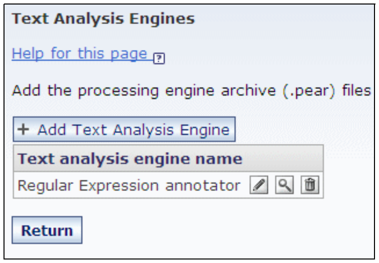
Path

○ Local path (the maximum file size is 64 MB)
File name:

● Index server path (the file can be larger than 64 MB)
File name:
/opt/IBM/es/packages/uima/apache-uima-regex/RegExAnnotator.pear

Figure 11-19 Adding a Text Analysis Engine

The window reflects the added engine (Figure 11-20).



Text Analysis Engines

Help for this page

Add the processing engine archive (.pear) files

+ Add Text Analysis Engine

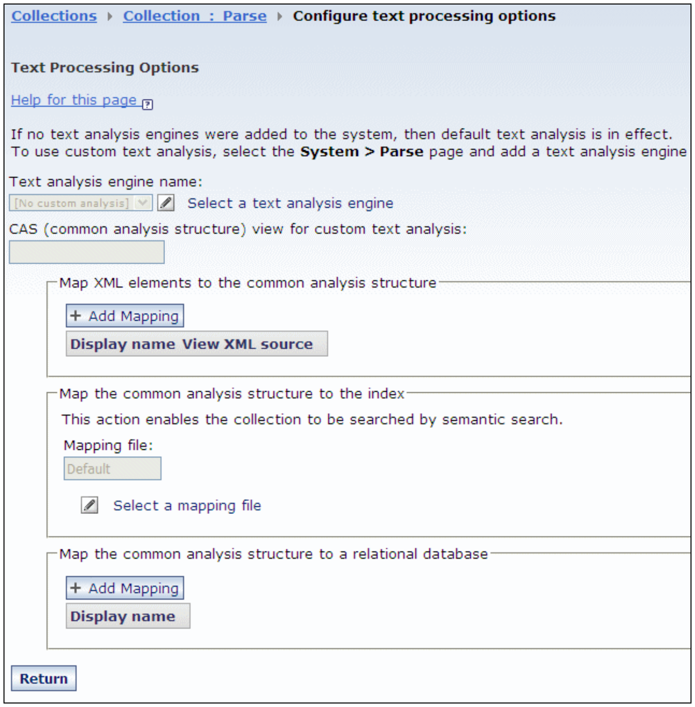
Text analysis engine name
Regular Expression annotator

Return

Figure 11-20 List of added text analysis engines

6. Select the **Collections** tab at the top of the administration console, and click the **Parse and Index** button of the collection with which you want to use a custom annotator.
7. Select **Text Analytics** → **Edit** → **Configure text processing option**.
8. In the Text Processing Options window (Figure 11-21), specify the configuration settings:

   a. Click **Select a text analysis engine**.



Collections ▸ Collection : Parse ▸ Configure text processing options

Text Processing Options

Help for this page

If no text analysis engines were added to the system, then default text analysis is in effect. To use custom text analysis, select the **System > Parse** page and add a text analysis engine

Text analysis engine name:

[No custom analysis] Select a text analysis engine

CAS (common analysis structure) view for custom text analysis:

Map XML elements to the common analysis structure

+ Add Mapping

Display name	View XML source

Map the common analysis structure to the index

This action enables the collection to be searched by semantic search.

Mapping file:

Default

Select a mapping file

Map the common analysis structure to a relational database

+ Add Mapping

Display name

Return

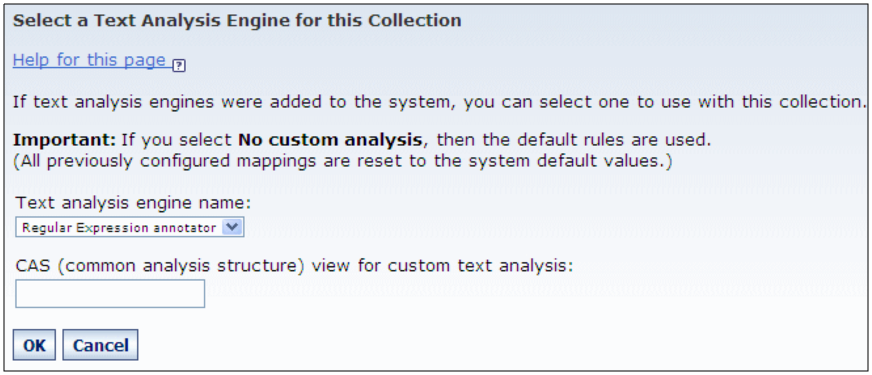
Figure 11-21 Configuring the text processing options

b. In the Select a Text Analysis Engine for this Collection window (Figure 11-22), complete these steps:

   i. Select the text analysis engine that you want to use with the collection.

   ii. Enter the CAS view name if necessary. In this figure, any name is not specified. When the name is specified, the Content Analytics document processor creates a separate CAS view to store UIMA annotations created by the custom analysis engine. It is used to avoid annotation conflicts. You have to enter an appropriate view name when using the PEAR file that is exported by LanguageWare Resource Workbench.

   iii. Click **OK**.



Select a Text Analysis Engine for this Collection

Help for this page

If text analysis engines were added to the system, you can select one to use with this collection.

Important: If you select **No custom analysis**, then the default rules are used.
(All previously configured mappings are reset to the system default values.)

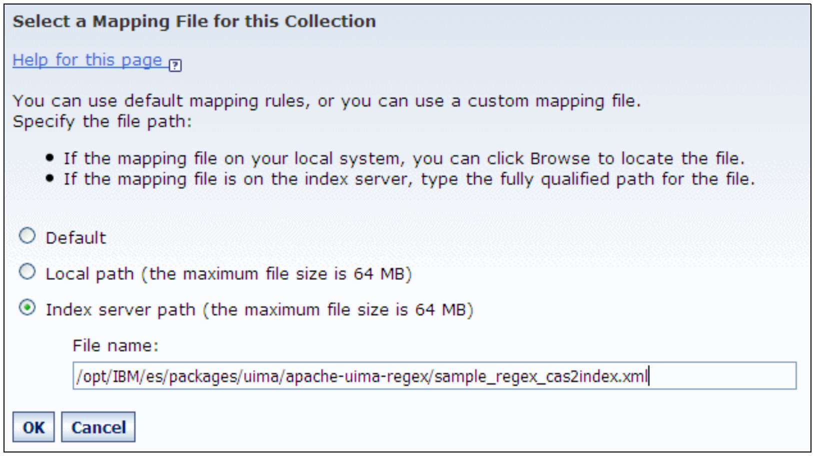
Text analysis engine name:
Regular Expression annotator

CAS (common analysis structure) view for custom text analysis:

Figure 11-22 Associating a text analysis engine with a collection

c. Back in the Text Processing Options panel (Figure 11-21 on page 471), click **Select a mapping file**.

d. In the Select a Mapping File for this Collection panel (Figure 11-23), specify the `sample_regex_cas2index.xml` file, and click **OK**.



Select a Mapping File for this Collection

Help for this page

You can use default mapping rules, or you can use a custom mapping file.
Specify the file path:

- If the mapping file on your local system, you can click Browse to locate the file.
- If the mapping file is on the index server, type the fully qualified path for the file.

○ Default

○ Local path (the maximum file size is 64 MB)

● Index server path (the maximum file size is 64 MB)

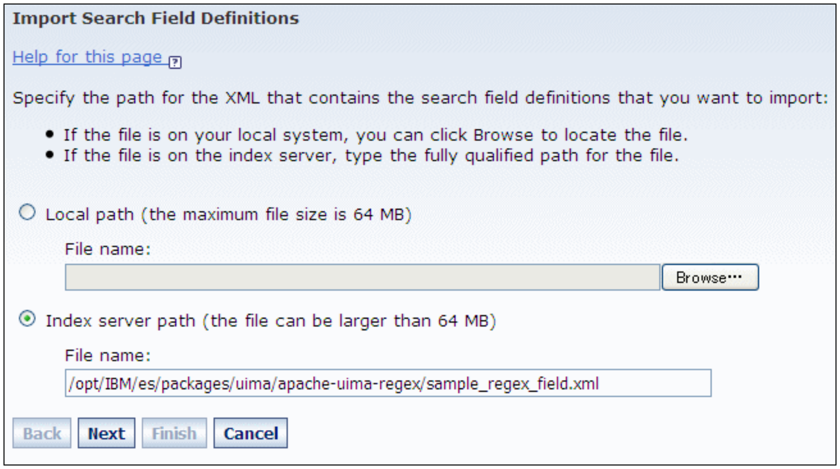
File name:

/opt/IBM/es/packages/uima/apache-uima-regex/sample_regex_cas2index.xml

Figure 11-23 Selecting a mapping file for this collection

e. When the window reflects the configurations, confirm them, and click **OK**.

9. Click **Configure search fields**.
10. When the window shows the search field definitions, click **Import Search Field**.
11. In the Import Search Field Definitions panel (Figure 11-24), specify the `sample_regex_field.xml` file, and click **Next**.



Import Search Field Definitions

Help for this page

Specify the path for the XML that contains the search field definitions that you want to import:

- If the file is on your local system, you can click Browse to locate the file.
- If the file is on the index server, type the fully qualified path for the file.

○ Local path (the maximum file size is 64 MB)

File name:

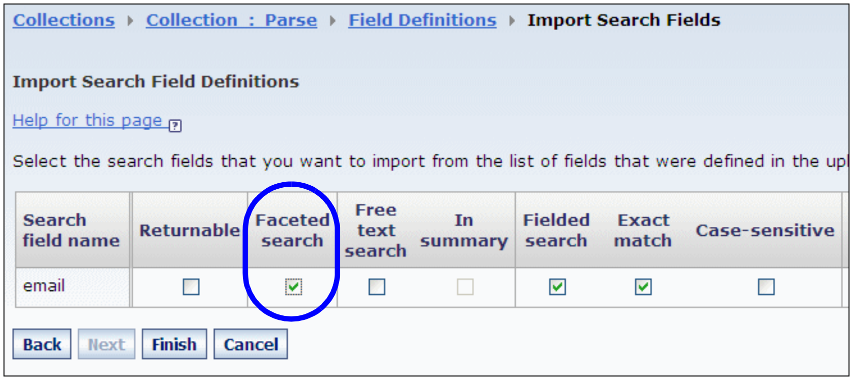
● Index server path (the file can be larger than 64 MB)

File name:

/opt/IBM/es/packages/uima/apache-uima-regex/sample_regex_field.xml

Figure 11-24 Selecting a search field definitions file to import

12. In the Import Search Field Definitions window (Figure 11-25), which shows the search field definition (`email`) that is imported from the `sample_regex_field.xml` file, select **Faceted search** for each imported search field. Click **Finish**.



Collections ▸ Collection : Parse ▸ Field Definitions ▸ **Import Search Fields**

Import Search Field Definitions

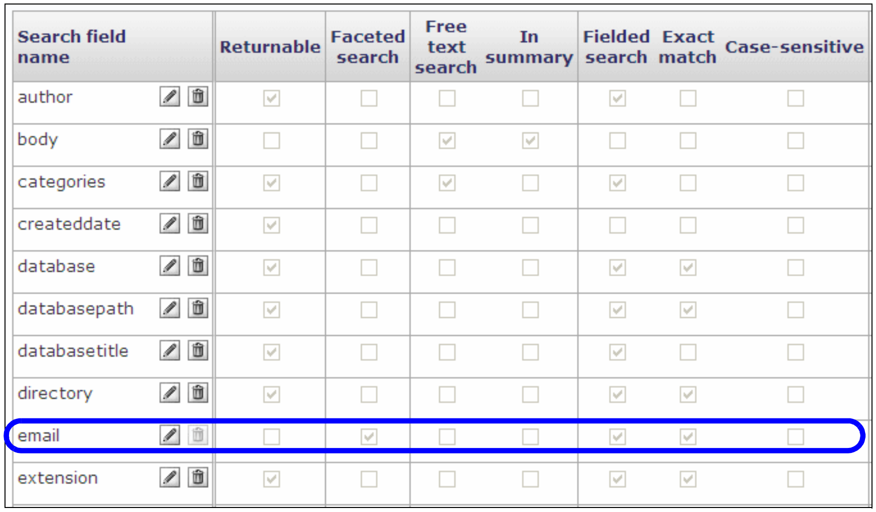
Help for this page

Select the search fields that you want to import from the list of fields that were defined in the upl

Search field name	Returnable	Faceted search	Free text search	In summary	Fielded search	Exact match	Case-sensitive
email	☐	☑	☐	☐	☑	☑	☐

Figure 11-25 Imported search fields definitions

13. In the definition table (Figure 11-26), verify that the imported field is displayed, and click **Return**.

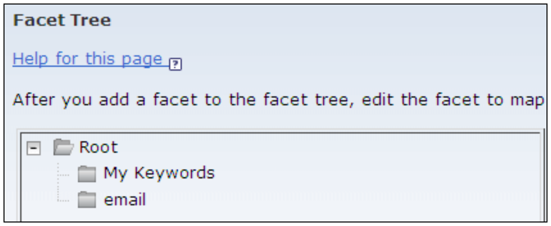


Search field name	Returnable	Faceted search	Free text search	In summary	Fielded search	Exact match	Case-sensitive
author	☑	☐	☐	☐	☑	☐	☐
body	☐	☐	☑	☑	☐	☐	☐
categories	☑	☐	☑	☐	☑	☐	☐
createddate	☑	☐	☐	☐	☐	☐	☐
database	☑	☐	☐	☐	☑	☑	☐
databasepath	☑	☐	☐	☐	☑	☑	☐
databasetitle	☑	☐	☐	☐	☑	☐	☐
directory	☑	☐	☐	☐	☑	☑	☐
email	☐	☑	☐	☐	☑	☑	☐
extension	☑	☐	☐	☐	☑	☑	☐

Figure 11-26 Search field definitions table

14. Click **Configure facets**.

15. Confirm that imported fields are added to the facet tree (Figure 11-27). If necessary, change the Facet name to a name that is easier to understand.

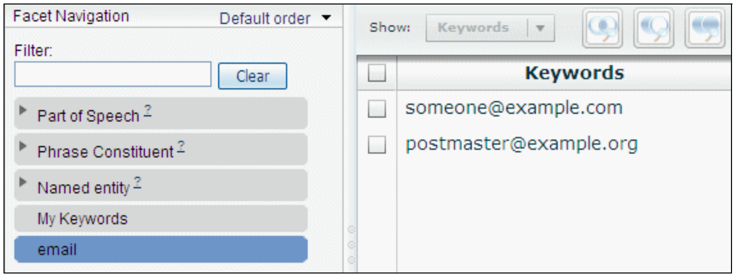

Facet Tree

Help for this page

After you add a facet to the facet tree, edit the facet to map

Figure 11-27 Facet tree showing added facets

After you complete these steps, redeploy the text analytics resources to use the custom annotators in your collection. Also rebuild an existing index to update the results in the text miner application. Figure 11-28 shows the email facet in the Facet view after rebuilding an index.



□	Keywords
□	someone@example.com
□	postmaster@example.org

Figure 11-28 Facet View showing the email facet that come from the Regular Expression annotator

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *custom text analysis integration*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

Use processing engine archive descriptor option: You might want to select the “Use processing engine archive descriptor” option when you deploy a PEAR file. In this case, you must specify the same data directory (`ES_NODE_ROOT`) on all document processing servers that you add to the system. When exporting a PEAR file directly to Content Analytics server using LanguageWare Resource Workbench, this option is always enabled.

11.3 Validation

You can validate your annotators by using the methods described in the following sections.

Validation hints and tips: Because an annotator can be validated only after it has been deployed, preserve a copy of the production environment. Also validate the changed annotators in the test environment before deploying the annotator in a real production environment.

11.3.1 Real-time NLP

This section provides information about the Real-time NLP capability as a stand-alone validation tool for the UIMA document processing pipeline in Content Analytics. The document processing pipeline of Content Analytics includes a series of annotators that you can configure. For example, you can update the dictionaries or the pattern rules and use more advanced annotators such as the Classification Module annotator or a custom built UIMA-compliant annotator. One validation methodology for these annotators is to use Real-time NLP.

Real-time NLP provides the capability to perform text analytics for a specified document that returns the same keywords that the text miner application might have shown. Real-time NLP also shows UIMA annotations in the CAS. Real-time NLP uses the same document pipeline defined for a given collection. It does not add any analysis results to the index. The benefit is that you can immediately check the analysis results without waiting for the index update.

Using Real-time NLP: To use Real-time NLP, you must enable the *Parse and Index* and *Search* components.

Testing the full UIMA pipeline using Real-time NLP

Annotators influence the behavior of Content Analytics by adding or changing a facet or by populating a search field. Adding user dictionaries and patterns also influences the behavior of Content Analytics. The easiest way to validate the success of changes reflected in facets is to use Real-time NLP. You can take advantage of Real-time NLP as explained in the following sections.

Using the Search REST API

Content Analytics provides the Search REST API that is a remote API set of the discovery functionality. You can easily use Real-time NLP from the Search REST

API through a browser. On machines where a search server has been deployed, the Real-time NLP REST API is available at the following address:

```
http://<Search Server Name:Port Number>/api/v10/analysis/text
```

Port Number is the search server port number that you specified during installation. By default, the port number is 8394.

You can use both HTTP GET and HTTP POST to call the Real-time NLP REST API. Example 11-1 is a sample HTML file to submit a document to a target collection.

Example 11-1 HTML form to submit a document to the Real-time NLP REST API

```
<html>
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
  </head>
  <body>
    <form method="post" enctype="multipart/form-data" action="http://<Host Name:Port
Number>/api/v10/analysis/text">
      <p><textarea rows="10" cols="80" name="text"></textarea></p>
      <p>Collection ID : <input type="text" name="collection" /></p>
      <p><input type="submit" value="Submit" /></p>
      <input type="hidden" name="output" value="export" />
    </form>
  </body>
</html>
```

After you submit the document, you see text analysis results as shown in Example 11-2.

Example 11-2 Response XML from the Real-time NLP REST API

```
<Document Id="resturi" Type="NORMAL">
  <Content Truncated="false" Encoded="false"><![CDATA[I could only find
11 cups in the 12-pack.]]></Content>
  <Metadata>
    <Fields>
      <Field Name="date">1288678851000</Field>
      <Field Name="$source">api</Field>
      <Field Name="$language">en</Field>
      <Field Name="$doctype">text/plain</Field>
      <Field Name="$charset">UTF-16</Field>
    </Fields>
    <Facets>
      <Facet>
```

```
        <Path>date</Path>
        <Path>2010</Path>
        <Path>11</Path>
        <Path>2</Path>
        <Path>15</Path>
      </Facet>
      <Facet Begin="8" End="12">
        <Path>_word</Path>
        <Path>adv</Path>
        <Keyword>only</Keyword>
      </Facet>
      <Facet Begin="13" End="17">
        <Path>_word</Path>
        <Path>verb</Path>
        <Keyword>find</Keyword>
      </Facet>
      <Facet Begin="18" End="20">
        <Path>_word</Path>
        <Path>num</Path>
        <Keyword>11</Keyword>
      </Facet>
      <Facet Begin="21" End="25">
        <Path>_word</Path>
        <Path>noun</Path>
        <Path>general</Path>
        <Keyword>cup</Keyword>
      </Facet>
      <Facet Begin="33" End="35">
        <Path>_word</Path>
        <Path>num</Path>
        <Keyword>12</Keyword>
      </Facet>
      <Facet Begin="36" End="40">
        <Path>_word</Path>
        <Path>noun</Path>
        <Path>general</Path>
        <Keyword>pack</Keyword>
      </Facet>
    </Facets>
  </Metadata>
</Document>
```

For details about the REST API, see the API documentation in the `ES_INSTALL_ROOT/docs/api/rest` directory.

Using the SIAPI

Content Analytics also provides the Search and Index API (SIAPI), which is a factory-based Java API. You can call the Real-time NLP using SIAPI. The sample `RealtimeNLPExample.java` application file is installed in the *`ES_INSTALL_ROOT`*`/samples/SIAPI` directory. You can run this application with Java development environments such as Eclipse.

To run this application, you must add the `siapi.jar` and `esapi.jar` files that are in the *`ES_INSTALL_ROOT`*`/lib` director to your class path. Then you must edit the following parameters:

- SERVER_HOSTNAME
- SERVER_PORT
- COLLECTION_ID

You must also specify a test document and the language of the document. There is no need to run a crawler to supply input to this sample program.

If the pipeline runs, the sample application returns all of the keywords that are extracted from the specified document. If a failure occurs, the sample application shows error messages that can help you diagnose the problem right away.

For details about the SIAPI API, see the API documentation in the `ES_INSTALL_ROOT/docs/api/siapi` directory.

When to use Real-time NLP

Consider using Real-time NLP in the following cases:

- You want to observe the impact of your custom dictionaries, patterns, or annotators in real time, which is the most typical case. Otherwise, to see analytics results in the text miner application, you must rebuild the index, which is an operation that takes time. With Real-time NLP, you can see keywords interactively for a given document.

  Example 11-3 shows the output of calling the Real-time NLP REST API after following the scenario in 7.2.1, “Scenario 1: Using a custom dictionary to discover package-related calls” on page 287. You can see the `$.myfacet.package` facet for the word “bottle.”

  Example 11-3 Output of Real-time NLP REST API after deploying custom dictionary

  ```
  <Facet Begin="18" End="31">
      <Path>_phrase</Path>
      <Path>noun_phrase</Path>
      <Path>adp_noun</Path>
      <Keyword>in ... bottle</Keyword>
  </Facet>
  ```

```
<Facet Begin="25" End="31">
    <Path>myfacet</Path>
    <Path>package</Path>
    <Keyword>bottle</Keyword>
</Facet>
<Facet Begin="25" End="31">
    <Path>_word</Path>
    <Path>noun</Path>
    <Path>general</Path>
    <Keyword>bottle</Keyword>
</Facet>
```

- You do not see the expected results nor any error messages:

  a. Check that you deployed the updated resources.

  b. Find documents that might contain the expected keywords in your collection. The query-based search can help you find the documents.

  c. Use Real-time NLP to see extracted keywords in real time. By using this method, you can check your dictionaries, rules, or annotators.

- You experience error messages when deploying a custom annotator, which are typically the result of the following issues:

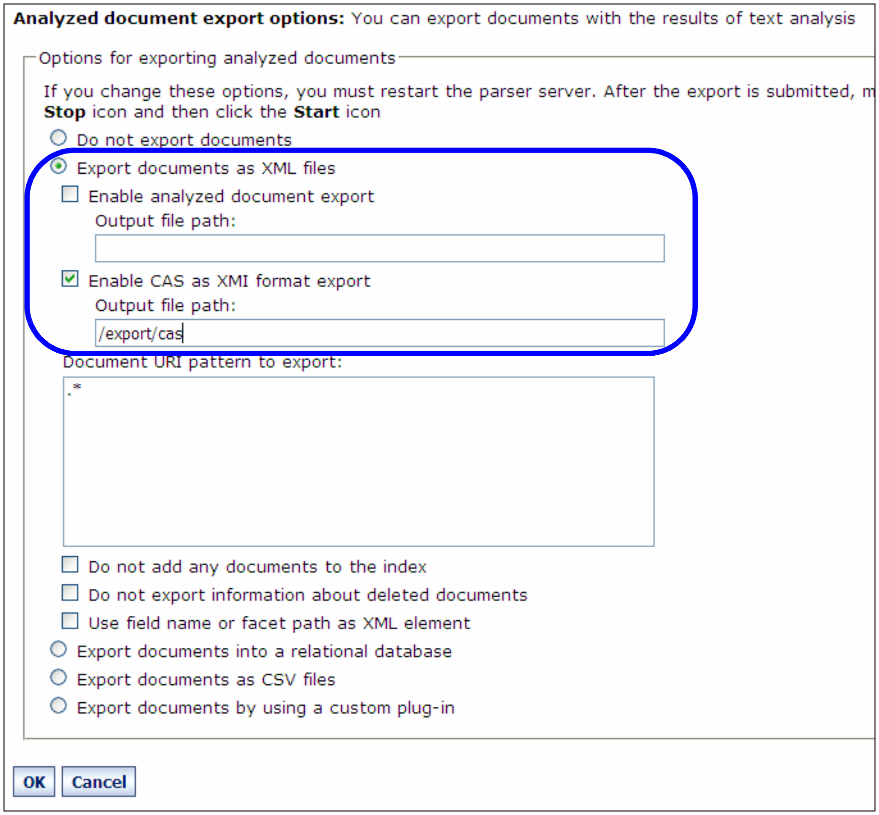
  – A UIMA annotator issue. The UIMA annotator (especially a custom annotator) does not work in Content Analytics.

  – An indexing issue. The UIMA annotator works as expected. However, the mapping between CAS and Content Analytics index is incorrect.

  In such cases, Real-time NLP helps to diagnose possible causes by showing error messages from the system or UIMA directly.

11.3.2 Advanced techniques

To investigate the problem in a collection deeply, you can use the **Configure options to export crawled or analyzed documents** option to export XMI files, which are the full serialized form of the CAS. Figure 11-29 on page 481 shows optional configurations to export the serialized CAS. For more information about exporting data, see Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387.



Analyzed document export options: You can export documents with the results of text analysis

Options for exporting analyzed documents

If you change these options, you must restart the parser server. After the export is submitted, m
Stop icon and then click the **Start** icon

○ Do not export documents

● Export documents as XML files

□ Enable analyzed document export

Output file path:

☑ Enable CAS as XMI format export

Output file path: /export/cas

Document URI pattern to export: .*

□ Do not add any documents to the index

□ Do not export information about deleted documents

□ Use field name or facet path as XML element

○ Export documents into a relational database

○ Export documents as CSV files

○ Export documents by using a custom plug-in

Figure 11-29 Enabling XMI export

Even though the CAS Visual Debugger (CVD) tool is not officially supported by Content Analytics, you can use it to visualize the CAS. This tool comes as part of the Apache UIMA package. For more information, see the UIMA documentation at the following web address:

http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/tools/tools.html#ugr.tools.cvd

CAS debugging: CAS debugging and troubleshooting is an advanced topic that requires a deep understanding of UIMA. IBM does not provide support for CAS investigation.

11.3.3 Summary of validation techniques

To summarize, this section provides validation techniques for the UIMA pipeline and custom annotators for your reference.

UIMA pipeline validation

UIMA pipeline validation includes the following techniques:

- Confirm the annotations by using the Facets view. Deploying the resources is required. Make sure to rebuild the index if the resources are updated.
- Use the dictionary candidates versus specific candidates. You can use this feature to test your new dictionary and switch between several words before deciding which are the best words to use. Go to the IBM Content Analytics Information Center at the following address. Then search on *custom user dictionaries for text analytics collections* to learn how to work with candidates in the dictionary and test them versus the already deployed ones:

  http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

- Reindex each time, and run a test query set.
- Import and export. You can use import and export on a per-document basis by using the **esadmin** command. For more information, go to the IBM Content Analytics Information Center at the following address, and search on *exporting and importing collection configurations*:

  http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Validation techniques for custom annotators

Validation for custom annotators includes the following techniques:

- Check the configuration, the names, and the logs.
- Use the LanguageWare Resource Workbench to see annotation results on the Content Analytics pipeline. See “Analyzing documents using a Content Analytics pipeline” on page 466.
- Use Real-time NLP to validate the annotators. See 11.3.1, “Real-time NLP” on page 476.
- Use the export capability to investigate low-level UIMA results in CAS. See 11.3.2, “Advanced techniques” on page 480.

12

IBM Content Assessment scenario

Organizations dealing with an explosion of unstructured content are facing increasing issues in organizing and decommissioning content. IBM leads the market in addressing this issue with Content Assessment.

This chapter provides details about the IBM Content Assessment offering of which IBM Content Analytics is a critical part of the equation. This chapter introduces its content analytical and classification capabilities that empower an organization to decide which content to decommission while preserving and using the content-centric portions with business value. It explains how to use analytics to intelligently unlock information, discover, explore and automatically classify business content, increasing productivity and reducing the manual effort involved.

This chapter includes the following sections:

- Content Assessment offering
- Overview of Content Assessment
- Content Assessment workflow
- Records management and email archiving
- Preferred practices

12.1 Content Assessment offering

In the normal course of business, you deal with huge amounts of unstructured information. How much of the unstructured information is relevant to the business or task at hand? The general estimate is that a large number, around 80% or more, of your data can be unnecessary. For example, it can be over-retained, irrelevant, redundant, or duplicated.

How can you know which data has business value and which data is unnecessary? What can you do to preserve only that part of the content that has business value, is risk-related, or requires life-cycle governance? What can you do to be prepared to quickly address legal compliance requests on collecting unorganized content?

In the context of Content Assessment, *content decommissioning* is defined as the action of deleting irrelevant business content. You assess the content of your data, decommission business irrelevant content, and collect information for further legal compliance requirements especially in the legal area of electronic discovery (eDiscovery). Consider the following approaches:

- Preserve *all* the content. This approach is expensive and can expose your company to legal issues in the future.
- Delete *all* the content. This approach represents an infringement to compliance regulations and can expose your company to legal issues in the future.
- Smartly identify, select, and preserve only what you need. This method is the best approach, but it is complex because you are required to deal with such a large amount of content.

Content Assessment empowers content decommissioning through exploration and insight. It helps you in the process of accessing, visualizing, and analyzing content across the organization to evaluate whether it is necessary to your business.

At the time of this writing, the following products are included in the Content Assessment offering and their functionality for the Content Assessment use case:

- Content Analytics V2.1
  - Crawl and analyze content from various enterprise content sources.
  - Allow exploration of content from these various sources.
  - Export meaningful subsets of documents.

- IBM Classification Module V8.7
  - Integrate into Content Analytics to enhance analytics by providing categorization and clustering.
  - Integrate into Content Collector for File Systems to enhance collection-time decision making for each document.
- IBM Content Collector V2.1.1
  - Gather the documents that are exported from Content Analytics into an IBM Enterprise Content Management (ECM) repository.

12.1.1 Concepts and terminology

To understand the Content Assessment offering, you must understand the following key concepts and terminology associated with the offering and technology:

Dynamic analysis A broad marketing term for the interactive exploration capabilities of Content Analytics. Companies are interested in content assessment simply to understand their information better so that they can improve their decisions about what to do with their content.

Dynamic collection The ability to identify and preserve files crawled by Content Analytics and collected using Content Collector within the context of an eDiscovery scenario. This concept addresses specific requests for a company to collect and preserve files related to a special case.

eDiscovery readiness

Within the context of content assessment, the state of being proactive about your information governance. Rather than being reactive, be proactive about the governance of your content. Use offerings and products, such as Content Assessment, Content Collector, Classification Module, and Enterprise Records, to proactively govern and control your information. In this state, you can be confident and ready to respond to legal inquiries.

Content decommissioning

Within the context of Content Assessment, the action of deleting irrelevant business content.

12.2 Overview of Content Assessment

Content Assessment assists you in discovering valuable business information that is buried beneath irrelevant, obsolete, and duplicate content. It also helps you preserve and prepare your content for efficient eDiscovery.

With Content Assessment, you can identify the necessary information and decommissions the unnecessary information by using the following steps:

1. Dynamically analyze what you have.

   Make rapid decisions about business value, relevance, and disposition.

2. Decommission content that is unnecessary.

   Save costs and reduce risk by eliminating obsolete, over-retained, duplicative, and irrelevant content and the infrastructure that supports it.

3. Preserve and use the content that matters.

   Collect valued content to manage, trust, and govern throughout its life span in an enterprise-grade ECM platform. Uncover new business value and insight by integrating with other content analytics solutions. Collect content in response to legal requests for information.

Through text analysis, Content Assessment gives you the tools to efficiently investigate the content and make informed decisions. You can use the analysis capabilities of Content Assessment to help you achieve the following goals:

- Aggregate. Gather data from multiple content sources and types by using the large variety of crawlers that Content Analytics brings into Content Assessment.
- Correlate. Use the deep analysis of content that surfaces trends, relationship patterns, concepts, and anomalous associations, which is a set of unique capabilities that Content Analytics brings to Content Assessment. The correlations rely on various additional advanced text analysis provided by the Unstructured Information Management Architecture (UIMA) pipeline. They include text classification and information extraction provided by Classification Module.
- Visualize. Content Analytics provides easy-to-use, feature-rich views to quickly dissect large corpus of content and derive insight from your content
- Explore. Content Analytics interactively investigates content with faceted navigation and drills down to surface new insights and understanding.

This section explains these Content Assessment capabilities through the following scenarios:

- Content decommissioning scenario
- Dynamically analyzing and collecting your content

12.2.1 Content decommissioning scenario

This section describes a scenario in which you must use Content Assessment to decommission your content. The goal of the scenario is to reduce costs and risk by retaining only the necessary content.

To illustrate the power and benefits of Content Assessment, this scenario highlights a big company, Fictitious Software Company A (referred to as *the Company*), that stores data and content for many years.

The Company is a software company that produces thousands of different products, with a significant number of releases. Over time, the Company's product releases change, and some of them are discontinued. In many cases, the product names are changing over time. The Company has stored several terabytes of document content over the last 20 years.

To be compliant with new regulation rules, the Company is required to store valuable business data for a certain period depending on the nature of the documents. The Company spends a lot of time and money in locating documents that the lawyers request to support a pending lawsuit.

The Company is aware that it has a huge amount of content irrelevant for the current business and for the compliance regulations. The Company wants to decommission the content in a responsible way, for example, so that the important business-related information is not disposed.

The Company decides to prepare itself more proactively for other potential lawsuits and to reduce the effort invested today in searching and making available the necessary documents to its lawyers.

The Company decides to acquire an ECM compliance suite of products to help them manage their large amount of content and reduce the risks that they face today with their unorganized content. Their problem then becomes how to decide what and how they must organize their data to be compliant and easier to manage.

To solve their problem, the Company uses the Content Assessment offering as follows:

1. Preclean the data set to remove documents that do not contain analyzable text. This pre-analysis can be performed through different methods:

   – Filter the files by their extensions and remove the file types that cannot be analyzed such as images. The remaining subset of documents has the potential for being text-analyzable.

   – Set up a *search collection* in Content Analytics to identify the nontext data that cannot be analyzed by using metadata information such as the file type, file size, last access date, version, and directory. After the nontext data is identified, remove these files from the data set.

2. With some of the unnecessary documents removed, create a text analytics collection to further decommission documents. When creating the text analytics collection, enable the following functionality:

   a. Select the **Add default facets for content assessment** check box to automatically add the following default facets that are displayed in the text miner application:

      - File Extension
      - File Size
      - Last Modified Date

      Default content assessment facets: After the collection is created, the default content assessment facets cannot be enabled.

   b. Under Advanced options (Figure 12-1 on page 489), complete the following selections for the collection:

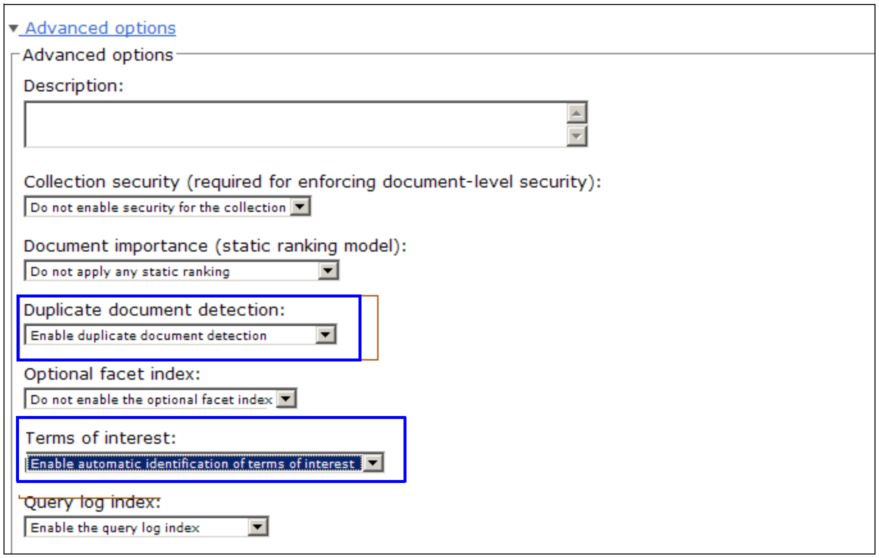
      i. In the Terms of interest field, select **Enable automatic identification of terms of interest**. With this function, you can identify relationships between nouns and nearby verbs and adverbs that exist in your content.

         For further information about terms of interest, see 8.1, “The power of dictionary-driven analytics” on page 322, and 8.2, “Terms of interest” on page 326.

      ii. In the Duplicate document detection field, select **Enable duplicate document detection**. This function identifies documents that might contain the same or nearly the same content. The duplicated documents are good candidates for being decommissioned and removed from your data set.

For details about duplicate document detection, go to the IBM Content Analytics Information Center at the following address, and search on *duplicate document detection*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp



▼ Advanced options

Advanced options

Description:

Collection security (required for enforcing document-level security): Do not enable security for the collection

Document importance (static ranking model): Do not apply any static ranking

Duplicate document detection: Enable duplicate document detection

Optional facet index: Do not enable the optional facet index

Terms of interest: Enable automatic identification of terms of interest

Query log index: Enable the query log index

Figure 12-1 Enabling the duplicate document detection and terms of interest functions

3. Create the crawler to your content source. By using the crawler capabilities of Content Analytics, the Company accesses various content sources, making the content available for inspection. Alternatively, instead of using crawlers, you can use the import from comma-separated value (CSV) file functionality to help import your filtered data into Content Analytics without having access to a repository. For more information, see 10.1, “Importing CSV files” on page 388.
4. Set up the clustering functionality for the collection. The clustering functionality is used as a method to gain quick insight into the content. For details about clustering, see 8.3, “Document clustering” on page 343.

5. Open the text miner application to gain a deeper view of your content:
   - Use the Facets, Deviations, and Trends views as explained in Chapter 6, “Text miner application: Views” on page 217, to better understand your content.
   - Enable the Named Entity Recognition annotator, and discover location names, persons names, and company names that can be found in the Company’s content.
   - Define custom facets and assign them values by using words lists.
   - Extract information by using patterns, such as social security numbers.

   In this scenario, the Company uses all the power of text analytics and the integration with Classification Module that Content Analytics offers. As a result, the Company gets a series of facets, such as the following examples, that reveal information about their content:

   - Intellectual Property
   - Legal Concepts
   - Sentiment Analysis
   - Products Analysis
   - Identify Personal and Corporate information (such as credit card numbers and various dates)
6. Inspect the content, and view it from different angles by using the text miner application. When the Company decides which data to preserve, the Company exports it by using the export capabilities to XML files of Content Analytics with the binary original content.
7. Reiterate this process from time to time in order to collect data to be further stored in an ECM repository.

   In this scenario, the Company has two business requirements:

   - Decommission content. The Company will invest a one-time effort to discover business-important data and preserve the data into the ECM repository. The Company will dispose of data from the old file shares or other old and obsolete data.
   - Periodically collect data and make it ready for legal eDiscovery if needed in the future. The Company needs to declare records according to the Company’s file plan that is defined during the Content Assessment process.

In Content Assessment, all activities can be inter-related, for example, content decommissioning, on-demand dynamic analysis for legal cases, periodic dynamic collection of content for eDiscovery readiness, and records declaration.

12.2.2 Dynamically analyzing and collecting your content

This section describes two scenarios in which you must use Content Assessment to dynamically analyze and collect your content.

In the first scenario, the *dynamic analysis* scenario, you must react rapidly to an eDiscovery request for legal compliance. You invest a lot of effort and resources in finding and retrieving the necessary documents. With proper eDiscovery enablement, you can reduce this effort significantly.

You start a content analysis task based on legal demand and when needed. The goal is to reduce eDiscovery costs and risk by performing eDiscovery collection across a broad range of enterprise sources as needed (in response to legal demands).

In the second scenario, the *dynamic collection* scenario, Content Analytics is used to collect content that is not currently being managed in an ECM repository. Content Analytics is also used to search and analyze the content according to an eDiscovery task order, or according to a broad search criteria, and export the content for ingestion. Content Collector is used to ingest the content and make it available for IBM eDiscovery tools, such as IBM eDiscovery Manager and IBM eDiscovery Analyzer.

The clustering and classification techniques of Classification Module are used in a support role to supplement the capabilities of both Content Analytics and Content Collector. The goal is to quickly investigate your data and address the legal requirement by providing the required data.

12.3 Content Assessment workflow

A major effort in planning for compliance readiness with Content Assessment is identifying what is important and has value for your business and disregarding what is unnecessary. Content Assessment helps you to explore your content and assists you in making informed decisions.

Figure 12-2 on page 492 illustrates a typical Content Assessment content decommissioning workflow:

1. Crawl, analyze, and index the content using Content Analytics.
2. Organize the content into meaningful hierarchies by using Classification Module, which can be deployed within Content Analytics.
3. Search and mine content by using the text miner application.

4. Export a resulting subset of documents including the associated metadata to a file share.
5. Use Content Collector to collect this exported content including the associated metadata into the ECM repository.
6. Dispose of the original content.

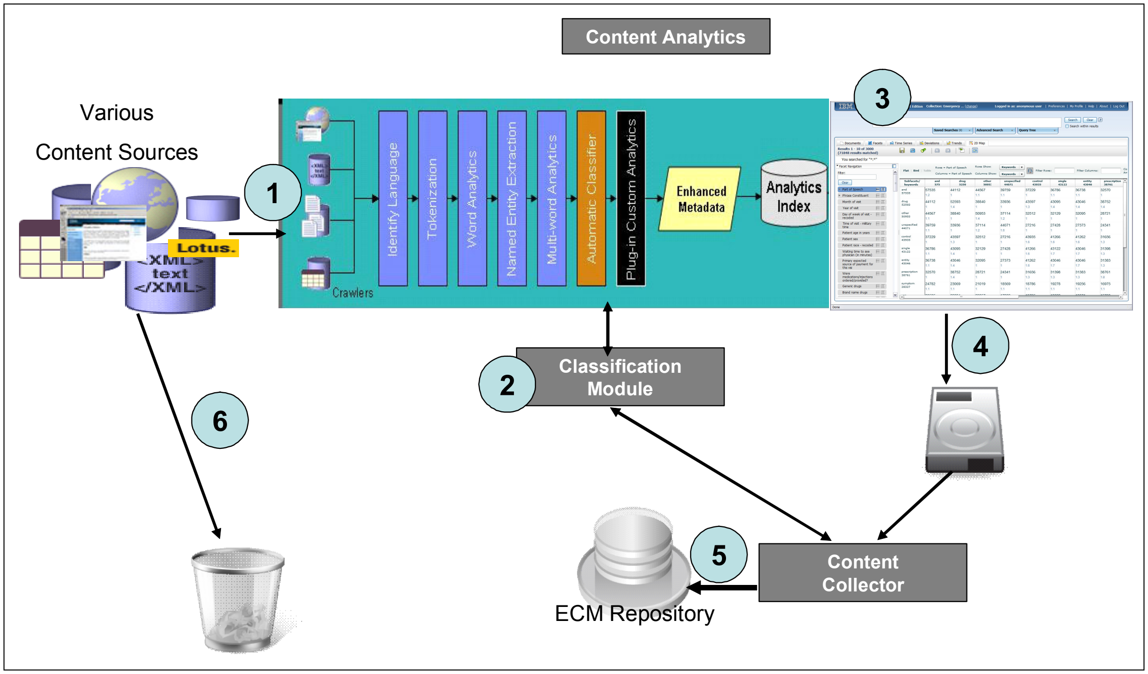


Figure 12-2 Typical content decommissioning workflow

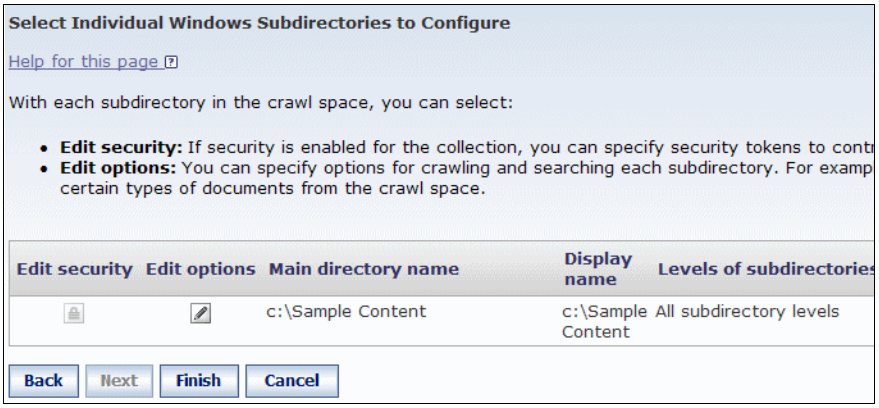
12.3.1 Decommissioning content

This section describes the procedures for content decommissioning by using Content Assessment. Content Analytics is used to gather and analyze content and to export the content with real business value. Content Collector is used to ingest the content that is exported by Content Analytics into an ECM repository. Classification Module is used to categorize the content for Content Analytics to provide additional facets for analytics. Classification Module is also used to categorize content for Content Collector to target the content to appropriate locations within the ECM repository or to assign additional metadata to the content. Next the unnecessary content is decommissioned.

Crawling, parsing, and indexing content with Content Assessment

To decommission content with Content Assessment, you must first crawl, parse, and index the content by using Content Analytics:

1. Identify the content sources that you want to explore. In this scenario for the Company, all documents are stored in the file system.
2. With the Content Analytics component of Content Assessment, crawl the content sources. Take a small sample for the content in the initial exploration step:
   a. Create a *text analytics collection*. In the collection name field, type `Fictitious Software Company A`.
   b. Configure the crawler. In this scenario, we configure a "Windows file system" crawler to crawl a sample of the Company's document stores on different files shares. Point the file crawler to the root location because Content Analytics will recursively crawl the subfolders. You can control the depth of the recursion. Figure 12-3 shows the Crawler Configuration Summary window.



Select Individual Windows Subdirectories to Configure

Help for this page

With each subdirectory in the crawl space, you can select:

- **Edit security:** If security is enabled for the collection, you can specify security tokens to contr
- **Edit options:** You can specify options for crawling and searching each subdirectory. For examp certain types of documents from the crawl space.

Edit security	Edit options	Main directory name	Display name	Levels of subdirectorie
🔒	✎	c:\Sample Content	c:\Sample Content	All subdirectory levels

Figure 12-3 Content Analytics Crawling Configuration Summary window

   c. Enable the terms of interest feature when creating the *text analytics* collection from the text analyzable content.

   With the terms of interest functionality, you can identify relationships between nouns and nearby verbs and adverbs in the text. To see how to enable the terms of interest function, see Figure 12-1 on page 489. For details about terms of interest, see 8.1, "The power of dictionary-driven analytics" on page 322, and 8.2, "Terms of interest" on page 326. Also, go to the IBM Content Analytics Information Center at the following address, and search on *terms of interest*:

   `http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

Export capability: If you plan to export, enable the export capability, and specify the full path for the destination folder. For more information about export, see Chapter 10, "Importing CSV files, exporting data, and performing deep inspection" on page 387.

3. Configure the Parse and Index options of Content Analytics.

   Before starting the crawler, set up Content Analytics to add some of the file system metadata as facets so that they can be used in the analysis tools. For Content Assessment, use the Named Entity Recognition annotator in the parsing pipeline.

   To configure the Parse and Index options, follow these steps:

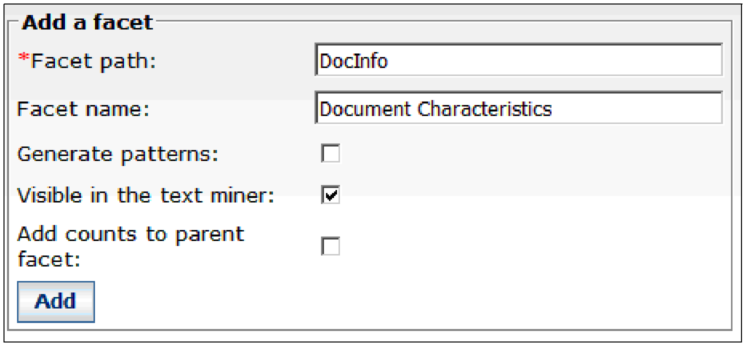
   a. Go to the **Parse and Index** tab of the Administration application.

   b. Turn on the Named Entity Recognition annotator if you think your data contains proper names, company names, or places. You can first inspect a small amount of your data to help you understand whether running this particular analysis can be beneficial. See 11.1.4, "Named Entity Recognition annotator" on page 452.

   c. Configure the Search fields. Go to the Field Definition page, and configure the existing fields or create new ones.

   d. Configure facets, and map them with the corresponding search fields. In this scenario, we add four facets and map each of them to a search field. In the Add a facet section (Figure 12-4), complete these steps:

      i. For Facet path, enter `DocInfo`.
      ii. For Facet name, enter `Document Characteristics`.
      iii. Select **Visible in the text miner**.
      iv. Click **Add**.



Add a facet

*Facet path: DocInfo

Facet name: Document Characteristics

Generate patterns: □

Visible in the text miner: ☑

Add counts to parent facet: □

Figure 12-4 Adding the Document Characteristics facet

v. Add four facets under the Document Characteristics facet: Size, Extension, Filename, and Directory (Figure 12-5). For information about how to add facets, see “Creating facets and mapping search fields to facets” on page 106.

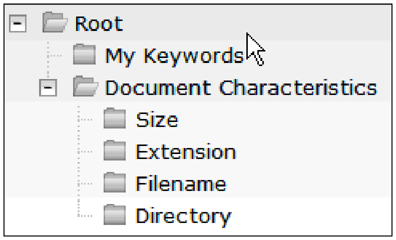


Figure 12-5 Facet tree with four document characteristics

e. Map the facets and the search fields as follows:

- Size (facet) to filesize (search field)
- Extension (facet) to extension (search field)
- Filename (facet) to filename (search field)
- Directory (facet) to directory (search field)

4. Run the crawler, parser, and indexer.

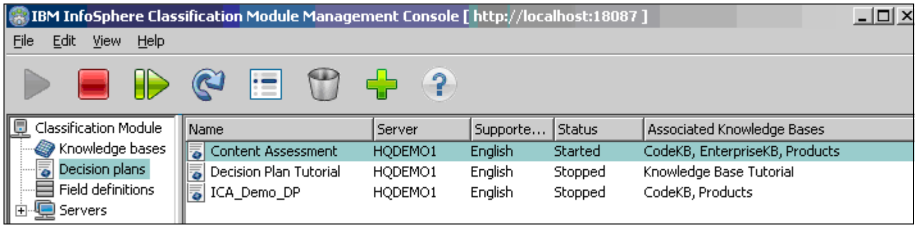
   Collecting content is the first step for content decommissioning and dynamic collection. For dynamic collection, you might want to set additional filters in the crawlers to look for specific types of content. Usually, the gathering of content for both use cases is the same.

5. Attempt to explore the data by using the text miner application of Content Analytics. See Chapter 5, “Text miner application: Basic features” on page 143, for information about application usage of text miner application.

Organizing content with Classification Module

You can use the Classification Module annotator to supplement native capabilities of Content Analytics by providing additional metadata for analytics and adding new facets. For this scenario, we use the following Classification Module knowledge bases and decision plans to initiate Content Assessment discovery:

- Content Assessment decision plan (Figure 12-6)

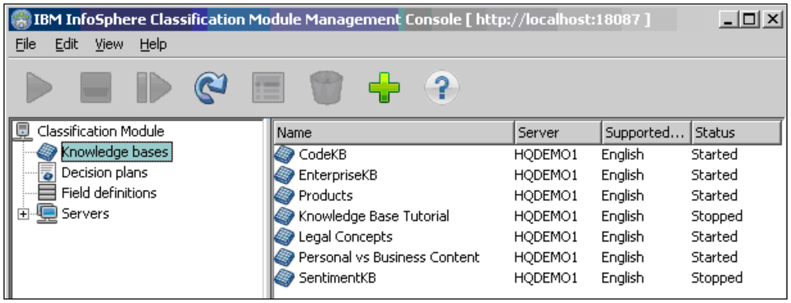


Name	Server	Supporte...	Status	Associated Knowledge Bases
Content Assessment	HQDEMO1	English	Started	CodeKB, EnterpriseKB, Products
Decision Plan Tutorial	HQDEMO1	English	Stopped	Knowledge Base Tutorial
ICA_Demo_DP	HQDEMO1	English	Stopped	CodeKB, Products

Figure 12-6 Decision plan in Classification Module

- Knowledge bases:
  – CodeKB
  – EnterpriseKB
  – Products
  – Legal Concepts
  – SentimentKB

Figure 12-7 shows the knowledge bases in Classification Module.



Name	Server	Supported...	Status
CodeKB	HQDEMO1	English	Started
EnterpriseKB	HQDEMO1	English	Started
Products	HQDEMO1	English	Started
Knowledge Base Tutorial	HQDEMO1	English	Stopped
Legal Concepts	HQDEMO1	English	Started
Personal vs Business Content	HQDEMO1	English	Started
SentimentKB	HQDEMO1	English	Stopped

Figure 12-7 Knowledge bases in Classification Module

Goal: From the content, discover facts that must be reviewed about the existing products. For example, analyze the documents to discover legal concepts that are relevant for the Company’s management. They attempt to identify personal and corporate information from within unstructured content such as credit card numbers and various dates.

To use the Classification Module annotator, follow these steps:

1. Verify that the Classification Module services are running. Make the Classification Module knowledge bases and decision plans available.
2. Enable the Classification Module integration. See Chapter 9, “Content analysis with IBM Classification Module” on page 357.
   a. Configure the connection of Content Analytics to Classification Module in the Content Analytics administration console.
      vi. Click the **Parse and Index** tab.
      vii. Click the **Edit** mode icon.
      viii.Click **Configure Classification Module**.
      ix. For the URL of the Classification Module server field, type the web service URL of your Classification Module, and click **Next**.
   b. For Decision plan, select **Content Assessment**.
   c. For Classification Module fields, select the following fields:
      - ContentCategory
      - IsCandidateForDeletion

      Each Classification Module field that you select on this page is mapped to a search field with the same name.
   d. Map the Content Analytics facets to the new search fields.

      Add new facets to present the text analytics information. For this example, we add patterns, such as Social Security Number (SSN).
   e. Add a facet under the Document Characteristics facet called `IsDeleteCandidate`. The label is “Deletion Candidate.”

      When a user selects the **Deletion Candidate** facet in the text miner application, the system selects documents that have been identified as deletion candidates by Classification Module.
   f. Under the Edit a Facet section, map the remaining fields that you added earlier. When you are finished, click **OK** to go back to the **Parse and Index** configuration tab.
3. Enable Classification Module in the parsing pipeline for Content Analytics. Click **My collection** to enable the Classification Module annotator. Click **OK**.
4. Index your data. Deploy the resources and reindex for the changes to the annotators and dictionaries to take effect.

Searching and mining content using the text miner application

Explore the content by using the text miner application of Content Analytics:

1. Start the Search Engine.
2. Using the text miner application, explore the content by running queries, and navigate through the facets. You can view the data from different aspects, such as entities extracted from the content (for example, people's names, grammatical phrases, or regular expressions such as SSNs).

   Go to the facet navigation pane to select which facets (or entities) you want to analyze. These facets can be extracted directly from the content itself or mapped to categories defined by IBM Classification Module, as described in Chapter 11, "Configuring annotators" on page 449.

   When exploring the data in this example, you see the following facets among others:

   – SSN
   – Phone Numbers
   – Credit Card Numbers

   For example, in the Facet Navigation pane (Figure 12-8), expand the **Document Characteristics** facet and expand the document content.

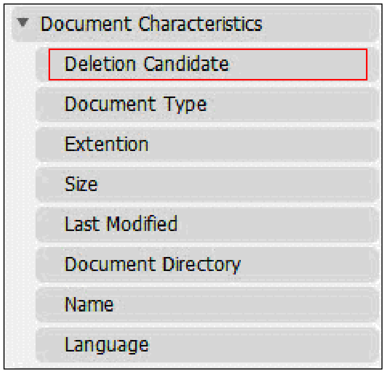
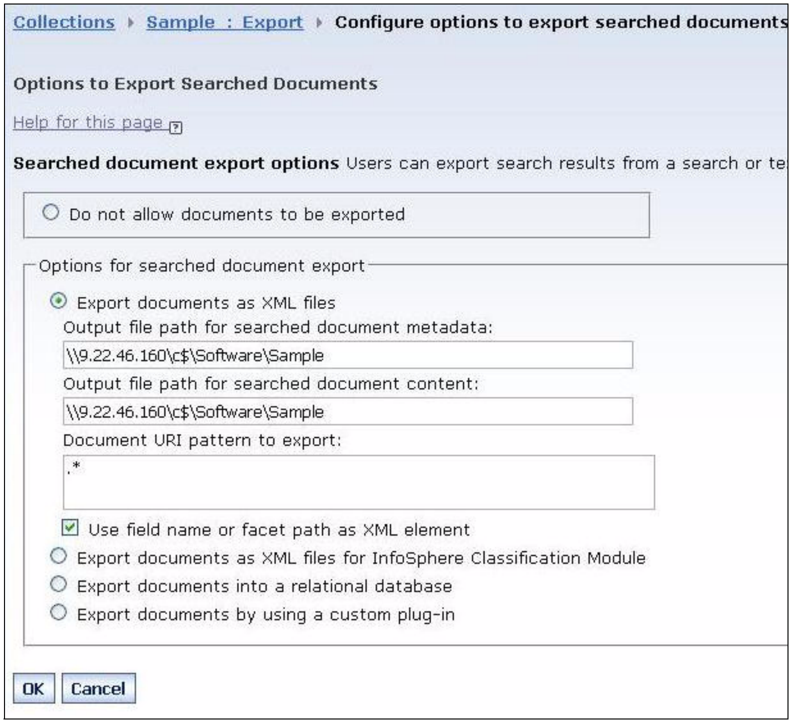


Figure 12-8 Expanding the Document Characteristics facet

   You see different credit cards values. This information is extracted by the Classification Module decision plan. You can review the respective documents. These documents have business value for the Company in this example. Therefore, we must preserve them.

3. Export both the binary files and the XML documents by using the Export functionality (Figure 12-9), which is explained in Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387.



Collections ▸ Sample : Export ▸ Configure options to export searched documents

Options to Export Searched Documents

Help for this page

Searched document export options Users can export search results from a search or te

○ Do not allow documents to be exported

Options for searched document export

● Export documents as XML files

Output file path for searched document metadata: \\9.22.46.160\c$\Software\Sample

Output file path for searched document content: \\9.22.46.160\c$\Software\Sample

Document URI pattern to export: .*

☑ Use field name or facet path as XML element

○ Export documents as XML files for InfoSphere Classification Module

○ Export documents into a relational database

○ Export documents by using a custom plug-in

Figure 12-9 Options to Export Searched Documents panel

> **Requirements for exporting data to a network file share:** To export data to a network file share that is accessible by Content Collector (read/write permission), you must configure Content Analytics. You must have Content Collector and FileNet P8 machines in an Active directory domain, and the Content Collector service must be running as a domain account.

4. Query your data set, and navigate through the facets to better understand your content.
5. Decide if you need to add more informational entities. Define more facets, and perform the following steps as necessary:
   a. Enhance the dictionary for the existing facets.
   b. Add more facets and associate them with the new dictionaries.

   c. Add more facets and define the new pattern matching rules. See 11.1.5, “Dictionary Lookup and Pattern Matcher annotators” on page 452.

   d. Enhance the pattern matching rules.

6. Use Classification Module to apply text classification and advanced metadata extraction.

7. Use the text miner application to identify related concepts. Refine the rules as needed.

8. Inspect a data sample with the Classification Module tools, and build a knowledge base and decision plan that incorporate all the feedback gathered by the subject matter expert analyst.

9. Review the results. Reiterate steps as necessary.

10. When you reveal content with interesting business value that you decide to preserve, export it.

Decommissioning the content

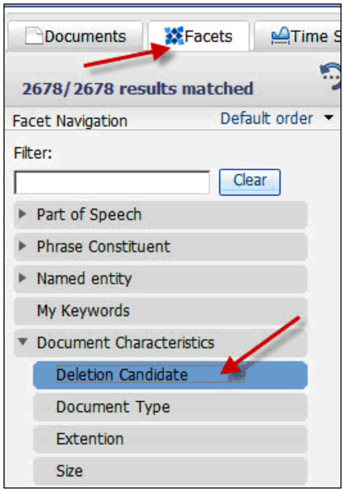
In this example, the people who manage the content have decided that one of the web content servers is full of unnecessary or redundant content. The only content that they want to keep are pages that deal with the *petrochemical* industry. The Company wants to copy this content to a newer server, and then the old content server can be taken out of service or decommissioned.

For this scenario, perform the following steps for content decommissioning:

1. Search the entire crawled content set. Use the IsDelete field to classify any document that is no longer necessary.

   The Classification Module was trained with a set of documents that are related to the petrochemical industry. For those documents that are not related to the petrochemical industry, the Classification Module sets the isDelete field of the documents to *Yes*. The isDelete field from the Classification Module is then mapped to the Deletion Candidate facet in Content Analytics.

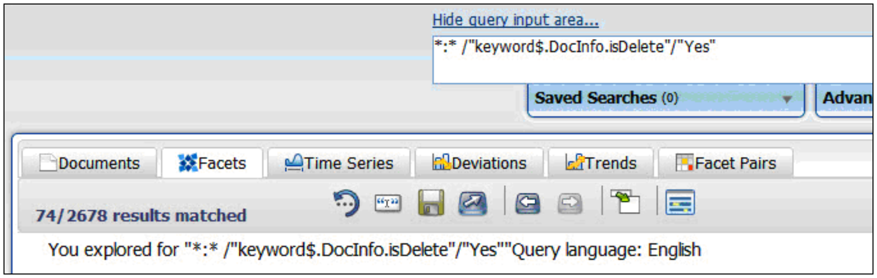
By checking on the documents under the Deletion Candidate in the Content Analytics view (Figure 12-10 on page 501), you see a list of documents that are candidates for deletion.



2678/2678 results matched

Figure 12-10 The Deletion Candidates facet

2. In the Content Analytics view (Figure 12-11), sort the content by Deletion Candidate criteria. This particular facet only has one category, Yes, meaning that the documents are related to the Petrochemical industry.



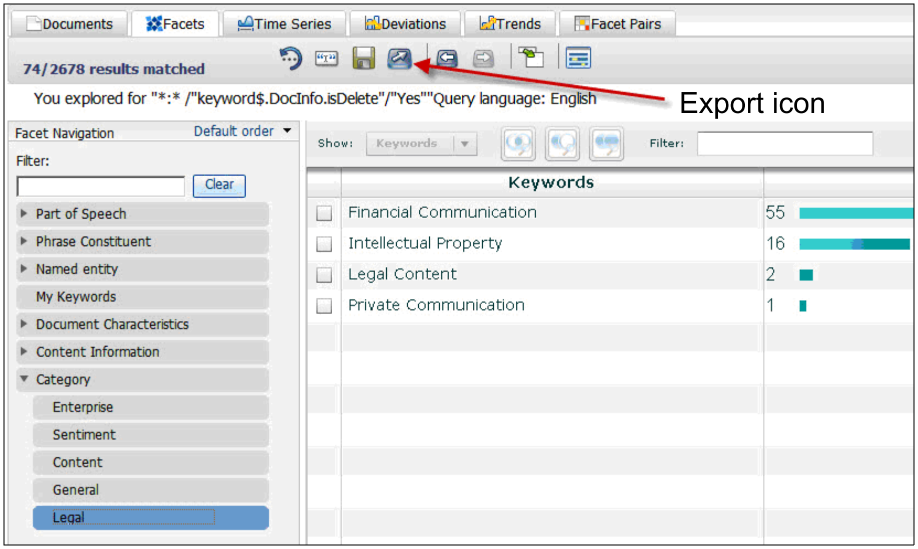
: /"keyword$.DocInfo.isDelete"/"Yes"

74/2678 results matched

You explored for "*:* /"keyword$.DocInfo.isDelete"/"Yes""Query language: English

Figure 12-11 Visualizing the results for isDelete documents

3. As shown in Figure 12-12, expand the **Category** facet, and click **Legal**. Then click the **Facets** tab.



74/2678 results matched

You explored for "*:* /"keyword$.DocInfo.isDelete"/"Yes""Query language: English

	Keywords	
□	Financial Communication	55
□	Intellectual Property	16
□	Legal Content	2
□	Private Communication	1

Figure 12-12 Viewing documents related to the Legal facet

Notice the four different legal categories. Classification Module has been trained to identify content that belongs to these categories. Therefore, you see a breakdown of all documents relating to the Petrochemical industry that are considered different types of legal documents.

You can use the **Export** icon (shown in Figure 12-12) to export this information to an external XML file. An application can then read this XML file and move the content associated with it to a different server, making it possible to decommission the existing server. In this task, do *not* click **Export**.

12.3.2 Performing dynamic analysis

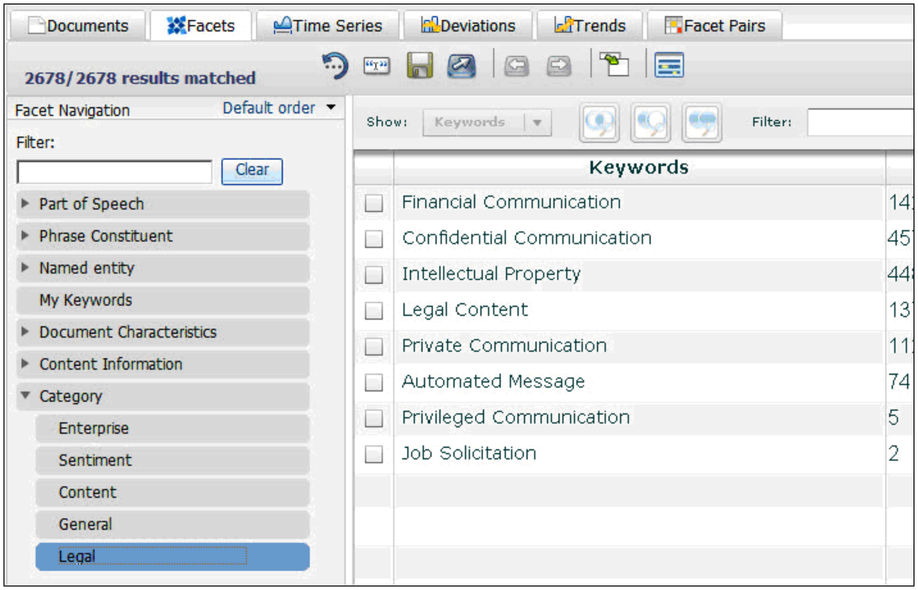
Dynamic analysis is a major use case of Content Assessment. Dynamic analysis refers to locating content across various repositories for eDiscovery.

Current eDiscovery tools require that content to be analyzed must be stored in a content repository such as *FileNet P8* or *IBM Content Manager.* This type of storage can be feasible for email messages. But, what if your content is spread across multiple repositories such as Sharepoint, web servers, Lotus Domino, or Documentum? What if the total volume of content is so large that you do not want to replicate it all in your FileNet P8 or IBM Content Manager repository?

The Company scenario: Assume that the Company is involved in litigation over the leaking of intellectual property information that occurred in its headquarters in early 2008. The Company's lawyers have asked you to find all content that might be applicable and to store it in a FileNet P8 repository where they can use eDiscovery Analyzer for further discovery.

To do dynamic collection, follow these steps:

1. Clear any previous queries, and start the text miner application.
2. Click the **Facets** tab (Figure 12-13). In the Facet Navigation pane, expand the **Category** facet, and select **Legal** documents.

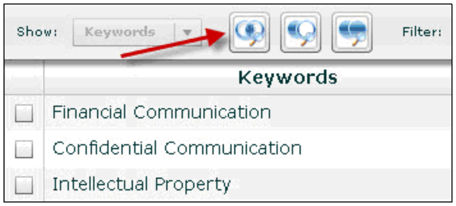


- Part of Speech
- Phrase Constituent
- Named entity
- My Keywords
- Document Characteristics
- Content Information
- Category
  - Enterprise
  - Sentiment
  - Content
  - General
  - Legal

	Keywords	
□	Financial Communication	14
□	Confidential Communication	45
□	Intellectual Property	44
□	Legal Content	13
□	Private Communication	11
□	Automated Message	74
□	Privileged Communication	5
□	Job Solicitation	2

Figure 12-13 Reviewing documents categorized under the Legal facet

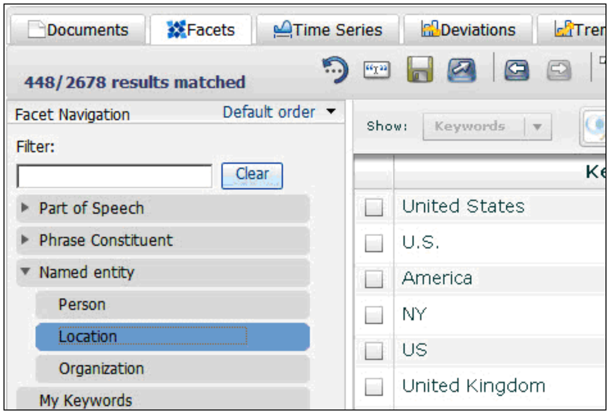
3. As shown in Figure 12-14, select the content related to Intellectual Property, and then click the **Add to search with Boolean ADD** icon.



	Keywords
□	Financial Communication
□	Confidential Communication
□	Intellectual Property

Figure 12-14 Choosing documents with the Intellectual Property keywords

4. Find the geographies that might be mentioned in the content for Intellectual Property. In the Facet Navigation pane (Figure 12-15), expand the **Named entity** facet, and select the **Location** facet.



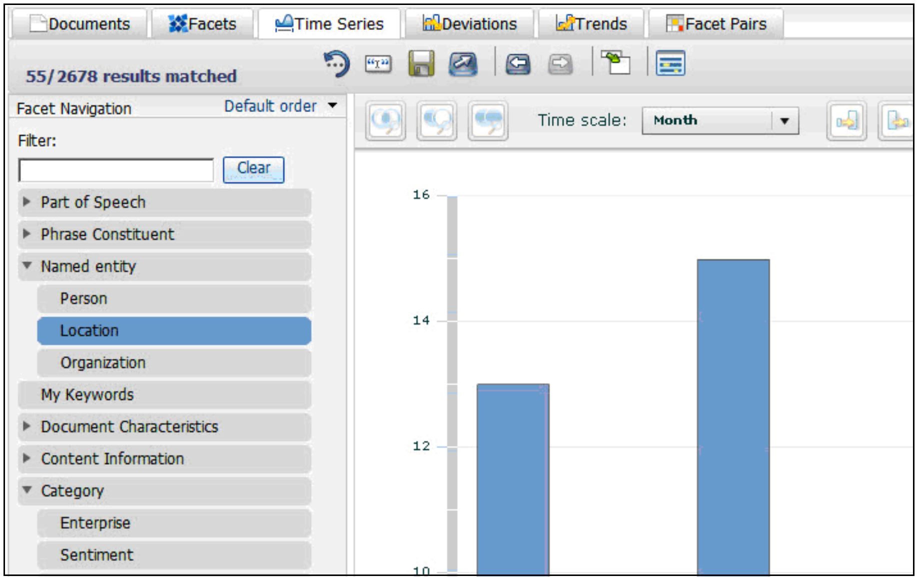
448/2678 results matched

□ United States
□ U.S.
□ America
□ NY
□ US
□ United Kingdom

Figure 12-15 Viewing documents within the Location facet

5. Focus on the documents related to the Company's main office. Select **NY** and click the **Add to search with Boolean ADD** icon.

6. Click the **Time Series** tab (Figure 12-16) to see how these documents are distributed by time. Change Time scale from Year to **Month**.



55/2678 results matched

Time scale: Month

Figure 12-16 Viewing documents for NY over time

7. Because you only want the content from the first quarter of 2008, draw a box around the bars for the first three months, as shown in Figure 12-17. The bars turn a darker color when they are selected. Click the **Add to search with Boolean ADD** icon.

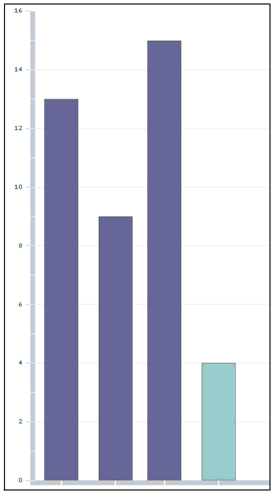


Figure 12-17 Focusing on documents from first quarter of 2008

Now you have the subset of the content that you want to store in the trusted repository for further analysis by the eDiscovery tools.

8. Click the **Export** icon.
9. From the drop-down list, select **Crawled content and parsed content with analysis results**. We want the crawled content so that we can collect and store it to FileNet P8. We also want the analysis results in case that information can be used to determine information such as record plans or storage locations. Leave the Schedulable radio button set to **No**.

   The Export utility is configured to store the exported content and metadata in the `C:\Export` directory.

10. Navigate to the `C:\Export` directory (Figure 12-18) using Windows Explorer. There are two subdirectories. A new directory with the current date and time has been created to store the exported metadata.

    Navigate the directory until you see a list of XML files.

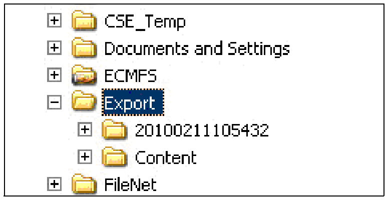
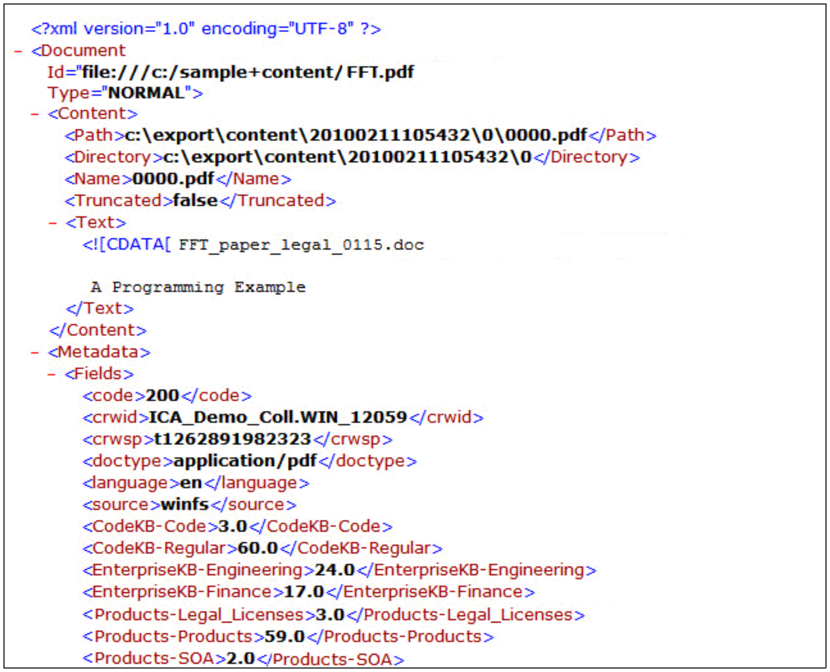


Figure 12-18 Exporting a directory for content and metadata

Figure 12-19 shows an example XML file and the type of data stored within it. The path to the original file tag called `Id` is where Content Collector collects the original file from.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <Document
    Id="file:///c:/sample+content/FFT.pdf
    Type="NORMAL">
  - <Content>
      <Path>c:\export\content\20100211105432\0\0000.pdf</Path>
      <Directory>c:\export\content\20100211105432\0</Directory>
      <Name>0000.pdf</Name>
      <Truncated>false</Truncated>
    - <Text>
        <![CDATA[ FFT_paper_legal_0115.doc

         A Programming Example
      </Text>
    </Content>
  - <Metadata>
    - <Fields>
        <code>200</code>
        <crwid>ICA_Demo_Coll.WIN_12059</crwid>
        <crwsp>t1262891982323</crwsp>
        <doctype>application/pdf</doctype>
        <language>en</language>
        <source>winfs</source>
        <CodeKB-Code>3.0</CodeKB-Code>
        <CodeKB-Regular>60.0</CodeKB-Regular>
        <EnterpriseKB-Engineering>24.0</EnterpriseKB-Engineering>
        <EnterpriseKB-Finance>17.0</EnterpriseKB-Finance>
        <Products-Legal_Licenses>3.0</Products-Legal_Licenses>
        <Products-Products>59.0</Products-Products>
        <Products-SOA>2.0</Products-SOA>
```

Figure 12-19 Content Analytics exported XML file

12.3.3 Preserving and using business data

This section explains how to preserve and use valuable business data. After inspecting the content and identifying the relevant business data, you need to preserve it and make it ready for further exploration and usage. You also need to preserve your valuable business data.

To assess the content and preserve the important data, follow these steps:

1. Search and mine the content:
   a. Crawl data from various data sources, which is especially important for the dynamic analysis scenario, by using the crawling function in Content Analytics.
   b. Narrow down specific subsets of documents that you want to preserve by using different search queries in Content Analytics.
   c. Identify the specific area in the data that you want to preserve and store for further eDiscovery operations by using the text miner application in Content Analytics.
2. Preserve the content by using the Export function in Content Analytics. Choose one of the following export options depending on the content inspection action you took previously:
   - **Search result documents** to export results of search queries that led to the content that you want to preserve and further store it in an ECM repository.
   - **Analyzed documents** to export documents along with their metadata, textual data, and any annotations added during the document processing pipeline. Examples include dictionary-based facets, classification-based facets, and pattern-based facets. The metadata and annotations can be further used when storing the documents in the ECM repository to populate the required field or to take advanced action, such as declaring them as records.
   - **Crawled documents**, which are documents that have been crawled but not parsed or analyzed. You can export metadata, binary content, or both. The binary document is the format that will be further stored into the ECM repository. The metadata, along with the analyzed data, is used to populate specific fields in the ECM repository.

See Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387, for more information.

You have now used Content Analytics to access your content by using crawlers, viewed content, and analyzed it by using text miner and Classification Module.

After deciding what data is important to preserve, you export the relevant data by using the export capabilities of Content Analytics.

During the process of refinement of the text miner resources and Classification Module resources, you can gradually build new dictionaries, rules, and new decision plans and knowledge bases with the Classification Module. When moving to the next step of using your data, you use Content Collector to collect the content and store it under FileNet P8 or IBM Content Manager CM8.

The content decommissioning scenario is primarily a one-time activity. However, the dynamic collection might need to be continuously run. In this scenario, the user has a specific request to collect and preserve files related to a specific case. In this case, after the initial dynamic analysis, the user sets up a Content Collector Task Route and invokes the Classification Module to collect data and store it into the ECM repository while analyzing it on the way. If you refined your Classification Module decision plans and knowledge bases, make them available for the data collection.

Configuring Content Collector to collect exported content

To configure Content Collector to collect exported content, complete the following tasks as explained in the sections that follow:

1. Preparing resources
2. Creating a task route in Content Collector for File Systems
3. Collecting the exported Content Analytics data

Preparing resources

To prepare for the resources, follow these steps:

1. Verify that the Content Collector services are up and running.
2. Use Content Collector for File Systems to upload the content into the FileNet P8 or Content Collector repositories. Content Collector for File Systems, with Classification Module, uploads the documents that are exported by using Content Analytics and organizes them based on the recommendations of Classification Module to the ECM content repository.
3. Use the Classification Module resources, such as decision plans and knowledge bases that you built and refined previously. See Chapter 9, “Content analysis with IBM Classification Module” on page 357, for information. The knowledge base or decision plan can be derived from the initial resources used during the information gathered by using the data inspection with Content Analytics. Alternatively, you can build new Classification Module resources as a result of the content understanding.

Creating a task route in Content Collector for File Systems

To create a task route in Content Collector for File Systems, follow these steps:

1. Start the Content Collector Configuration Manager (Figure 12-20).

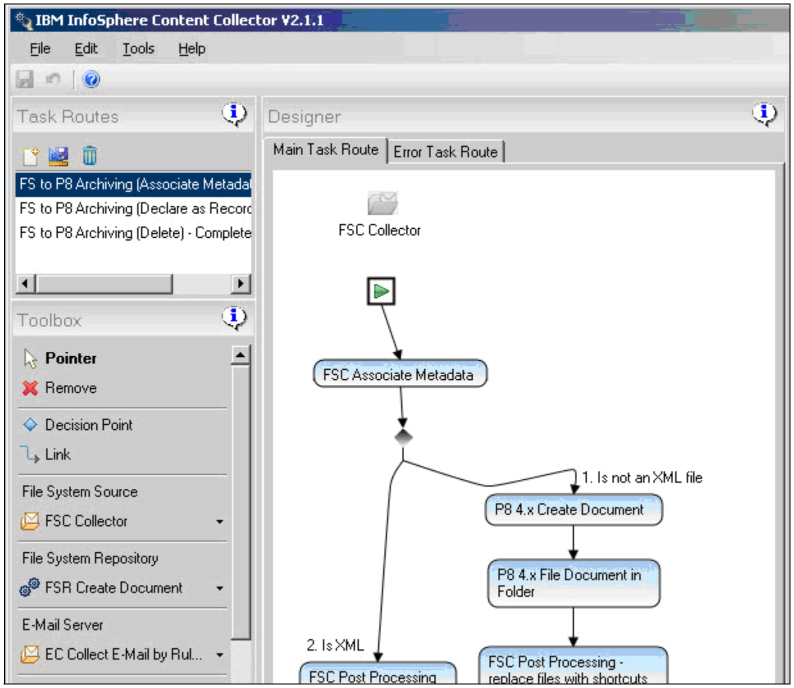


Figure 12-20 Content Collector Configuration Manager

2. Define the file system metadata that is used to help collect content. Use the elements in the XML files (exported from Content Analytics) to direct Content Collector where to find the files it needs to collect:

   a. Go to the Navigation panel (lower-left corner of the Content Collector Administration GUI) and select **Metadata and Lists** (Figure 12-21).

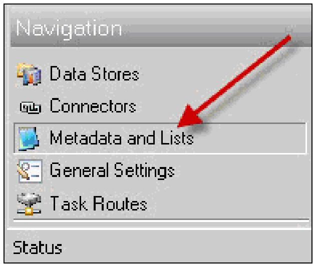
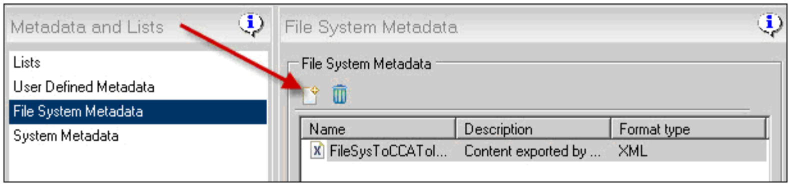


Figure 12-21 Selecting Metadata and Lists

b. In the left panel, select **File System Metadata**, and click **Add** to add the new file system metadata (Figure 12-22).



Name	Description	Format type
FileSysToCCATol...	Content exported by ...	XML

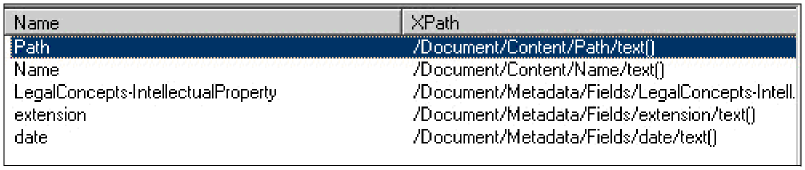
Figure 12-22 Adding the new file system metadata

c. Type a name for the new file system metadata, such as `Customer Metadata List for Content Assessment`.

d. Change Format Type to **XML**.

e. Select the XML elements that you want to use. Click the **Wizard** icon to start the XML file wizard.

f. Follow the Content Collector V2.1.1 documentation for defining the XML fields mapping. Figure 12-23 shows an example of such a mapping. Save your metadata mapping.



Name	XPath
Path	/Document/Content/Path/text()
Name	/Document/Content/Name/text()
LegalConcepts-IntellectualProperty	/Document/Metadata/Fields/LegalConcepts-Intell.
extension	/Document/Metadata/Fields/extension/text()
date	/Document/Metadata/Fields/date/text()

Figure 12-23 XML field mapping

3. Create the task route. You can create a task route from scratch or use one of the existing sample task routes as a starting point.

   For this example, we follow these steps:

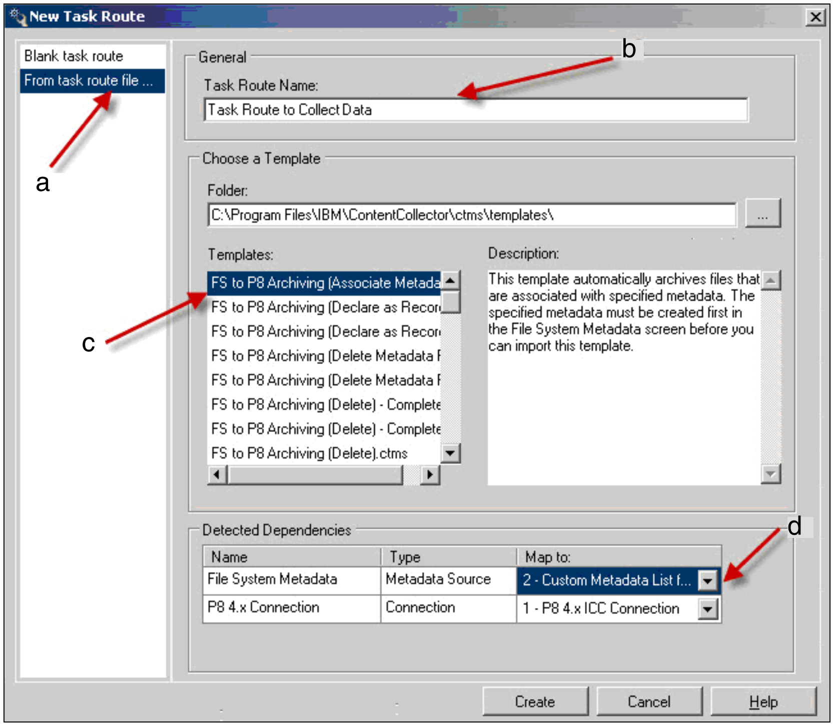
   a. From the New Task Route window, select **From Task Route file**.

   b. In the Task Route Name field, type `Task Route to Collect Data`.

   c. From the Template list, select **FS to P8 Archiving (Associate Metadata) - Complete** as a starting point.

   d. From the Detected Dependencies section, select the metadata mapping that you created for the File System Metadata mapping. In this case, for the FileNet P8 4.x connection, we select **P8 4.x ICC Connection**. If you do not have a valid FileNet P8 4.x connection to select, follow the Content Collector documentation instructions to fix this problem.

Figure 12-24 shows the configuration. Notice that the letters in the figure correspond to the substeps for step 3.



Task Route Name: Task Route to Collect Data

Folder: C:\Program Files\IBM\ContentCollector\ctms\templates\

Templates:

- FS to P8 Archiving (Associate Metada
- FS to P8 Archiving (Declare as Recor
- FS to P8 Archiving (Declare as Recor
- FS to P8 Archiving (Delete Metadata F
- FS to P8 Archiving (Delete Metadata F
- FS to P8 Archiving (Delete) - Complete
- FS to P8 Archiving (Delete) - Complete
- FS to P8 Archiving (Delete).ctms

Description: This template automatically archives files that are associated with specified metadata. The specified metadata must be created first in the File System Metadata screen before you can import this template.

Name	Type	Map to:
File System Metadata	Metadata Source	2 - Custom Metadata List f...
P8 4.x Connection	Connection	1 - P8 4.x ICC Connection

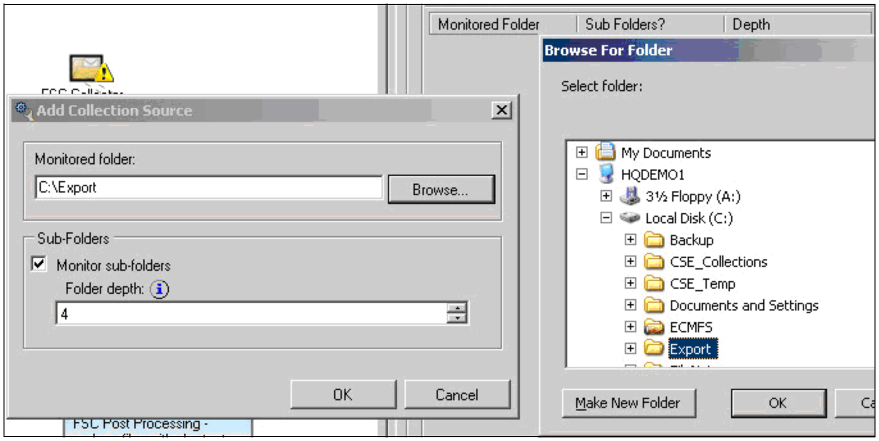
Figure 12-24 Defining a new task route

4. Modify the task route:

   a. In the Designer pane, click the **FSC Collector** icon. This icon represents the part of the file system from which you will collect content.

   b. Select the **Collection Sources** tab, and click **Add**.

c. Browse to the `c:\Export` directory, and select **Monitor sub-folders**. For the Folder depth field, type `4`, indicating for the system to monitor at least four levels deep for the subfolders (Figure 12-25). Click **OK**.


Add Collection Source

Monitored folder: C:\Export

Sub-Folders

☑ Monitor sub-folders

Folder depth: 4

Figure 12-25 Adding a collection source folder and specifying subfolder depth

5. Define the tasks in the Task Route:

a. Add a decision point.

b. Define the **FSC Associate Metadata** task.

c. Define a decision point, and connect the FSC Associate Metadata task to the P8 4.x Create Document task (Figure 12-26).

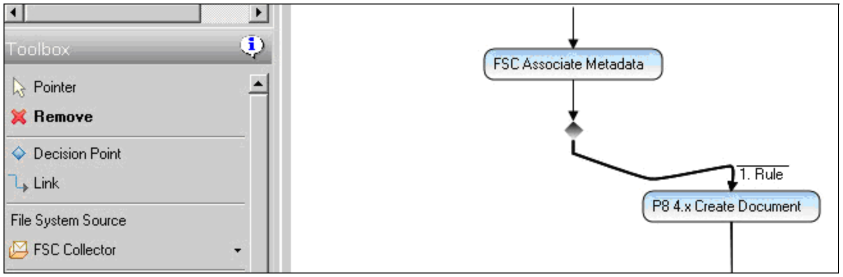
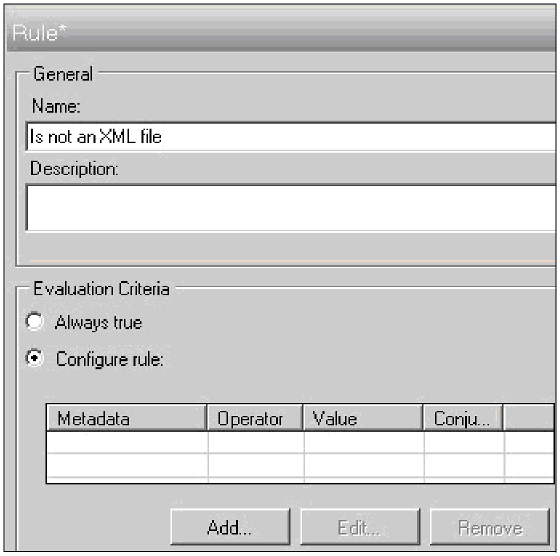


Figure 12-26 The FSC Associate Metadata task

d. Click the part of the link that connects the decision point to the P8 4.x Create the Document task. In the Rule definition panel (Figure 12-27), in the Name field, type the name of the rule `Is not an XML file`. Under Evaluation Criteria, select **Configure Rule**, and then click **Add**.



Rule*

General

Name: Is not an XML file

Description:

Evaluation Criteria

○ Always true

● Configure rule:

Metadata	Operator	Value	Conju...

Figure 12-27 Defining the 'Is not an XML file' rule

e. In the Edit Condition Clause window (Figure 12-28), complete these steps:

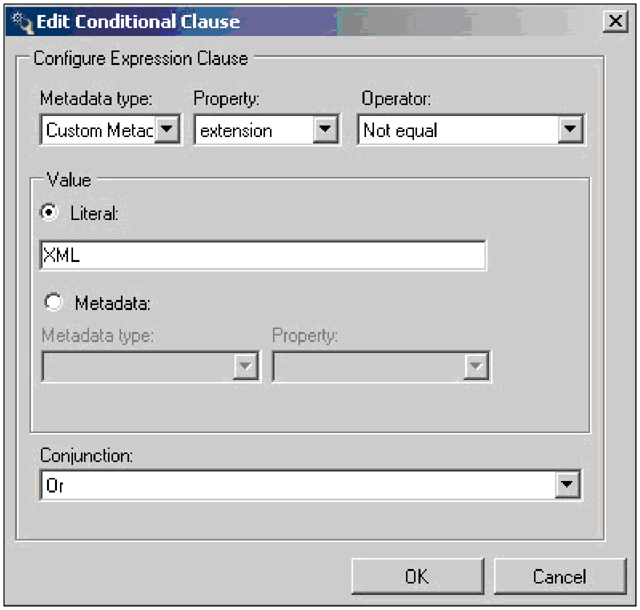
   i. In the Metadata Type field, select **Custom Metadata List for Content Assessment**.

   ii. For the Property field, select **extension**.

   iii. For the Operator field, select **Not equal**.

   iv. For the Value field, enter `XML`.

   v. Click **OK**.



Edit Conditional Clause

Configure Expression Clause

Metadata type: Custom Metac

Property: extension

Operator: Not equal

Value

● Literal:

XML

○ Metadata:

Metadata type:

Property:

Conjunction: Or

Figure 12-28 Editing the conditional clause for the rule

f. Add the FSC Post Processing task:

   i. Select **FSC Post Processing** from the list of possible FSC Collector tasks in the toolbar.

   ii. Drag the task to the far left side of the Designer window.

   iii. Right-click the decision point, and select **Add Rule**. A second link is displayed, called “2. Rule,” that comes from the decision point.

   iv. Click the **2. Rule** link, and drag it down so that it connects to the FSC Post Processing task that you just dragged to the Designer.

   Figure 12-29 shows the changes.

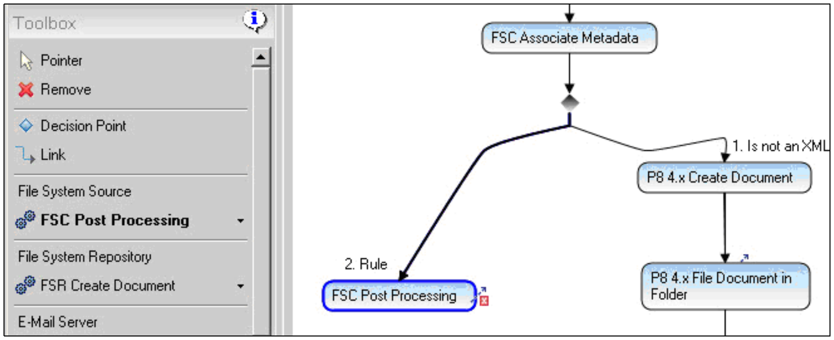


Figure 12-29 Adding the FSC Post Processing task

g. Click the **2. Rule** link (shown in Figure 12-29) that connects the decision point to the FSC Post Processing task. From the Rule pane on the right, change Rule name to `Is an XML file`. Select **Configure rule**, and click **Add**.

h. In the Edit Conditional Clause window (Figure 12-30), complete these steps:

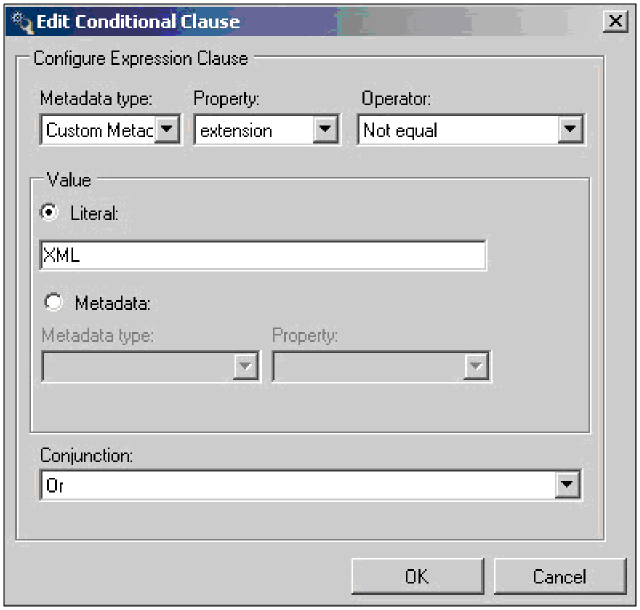
   i. In the Metadata field, select **Custom Metadata List for Content Assessment**.

   ii. For the Property field, select **extension**.

   iii. For the operator, select **Equal**.

   iv. For the Value field, select **Literal**, and enter `XML`.

   v. Click **OK**.



Edit Conditional Clause

Configure Expression Clause

Metadata type: Custom Metac

Property: extension

Operator: Not equal

Value

● Literal: XML

○ Metadata:

Metadata type:

Property:

Conjunction: Or

Figure 12-30 Adding the 'Is XML' rule

i. Select the **FSC Post Processing** task that is part of the "Is an XML file" rule, and edit the task:

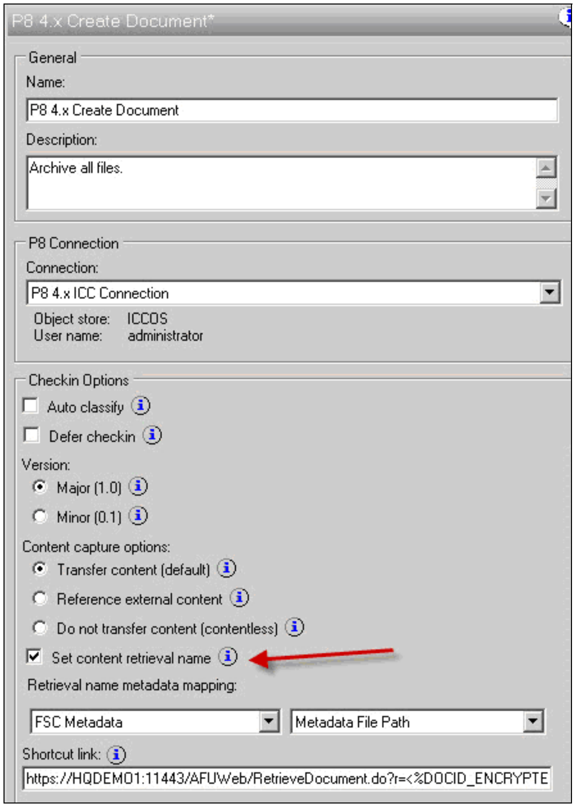
   i. Select **Delete File** from the Post Processing Options on the right side.

   ii. Clear the **Replace the file with a shortcut** check box.

j. Select the **P8 4.x Create Document** task that is part of the “Is not an XML file” rule. Then edit the task as shown in the P8 4.x Create Document window (Figure 12-31):

   i. Select the **Set content retrieval name** check box option.

   ii. In the Retrieval name metadata mapping field, select **FSC Metadata** from the left box, and select **Metadata file path** from the right box.



P8 4.x Create Document*

General

Name: P8 4.x Create Document

Description: Archive all files.

P8 Connection

Connection: P8 4.x ICC Connection

Object store: ICCOS
User name: administrator

Checkin Options

□ Auto classify
□ Defer checkin

Version:
● Major (1.0)
○ Minor (0.1)

Content capture options:
● Transfer content (default)
○ Reference external content
○ Do not transfer content (contentless)

☑ Set content retrieval name

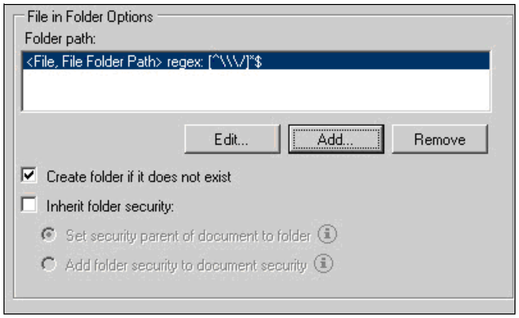
Retrieval name metadata mapping: FSC Metadata | Metadata File Path

Shortcut link: https://HQDEMO1:11443/AFUWeb/RetrieveDocument.do?r=<%DOCID_ENCRYPTE

Figure 12-31 Content retrieval name metadata mapping

k. Store the collected content in a FileNet P8 directory by updating the FileNet P8 4.x File Document in Folder task:

   i. Select the **FileNet P8 4.x File Document in Folder** task. Select the **Create folder if does not exists** option, and click **Add** (Figure 12-32).



File in Folder Options

Folder path:

<File, File Folder Path> regex: [^\\\/]*$

☑ Create folder if it does not exist

☐ Inherit folder security:

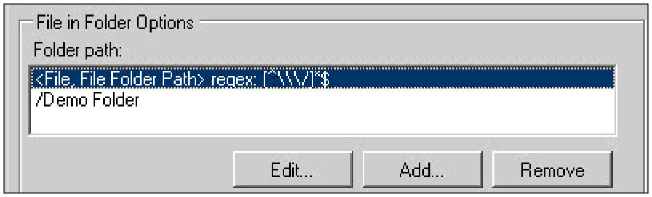
● Set security parent of document to folder

○ Add folder security to document security

Figure 12-32 Selecting the folder path to store content

   ii. Select **Demo Folder**, and click **OK**.

   iii. Under File in Folder Options (Figure 12-33), select the sample folder options definition, and click **Remove** to remove it.



File in Folder Options

Folder path:

<File, File Folder Path> regex: [^\\\/]*$

/Demo Folder

Figure 12-33 Selecting folders to store file content

l. Configure the audit setting for the FSC Post Processing task:

   i. Click the **Audit Log** icon that is connected to the task.

   ii. For auditing, select **Customer Metadata Lists for Content Assessment Lab**, **FSC Metadata**, and **P8 4.x Create Document**.

   iii. Right-click the **FSC Post Processing** task that is part of the "Is an XML File" rule (on the left), and select **Add Link Out**.

   iv. Click the newly added link, and connect it to the Audit Log task that is previously defined (Figure 12-34).

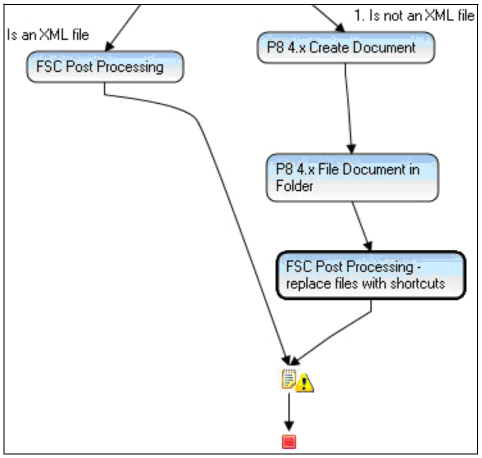

Is an XML file
1. Is not an XML file
FSC Post Processing
P8 4.x Create Document
P8 4.x File Document in Folder
FSC Post Processing - replace files with shortcuts

Figure 12-34 Linking the Audit Log task to the task route

6. Optional: To complete the Task Route, use Classification Module for additional classification.

   Figure 12-35 shows an example of a completed Task Route that you can use for Content Assessment, especially for the dynamic analysis and dynamic collection scenario.

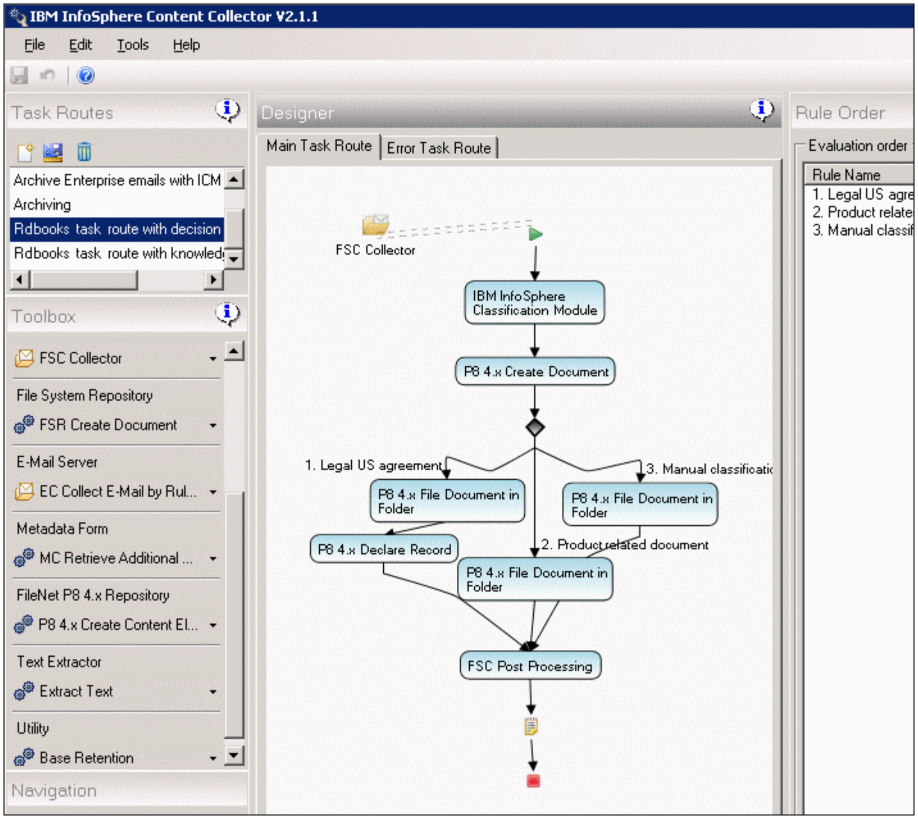
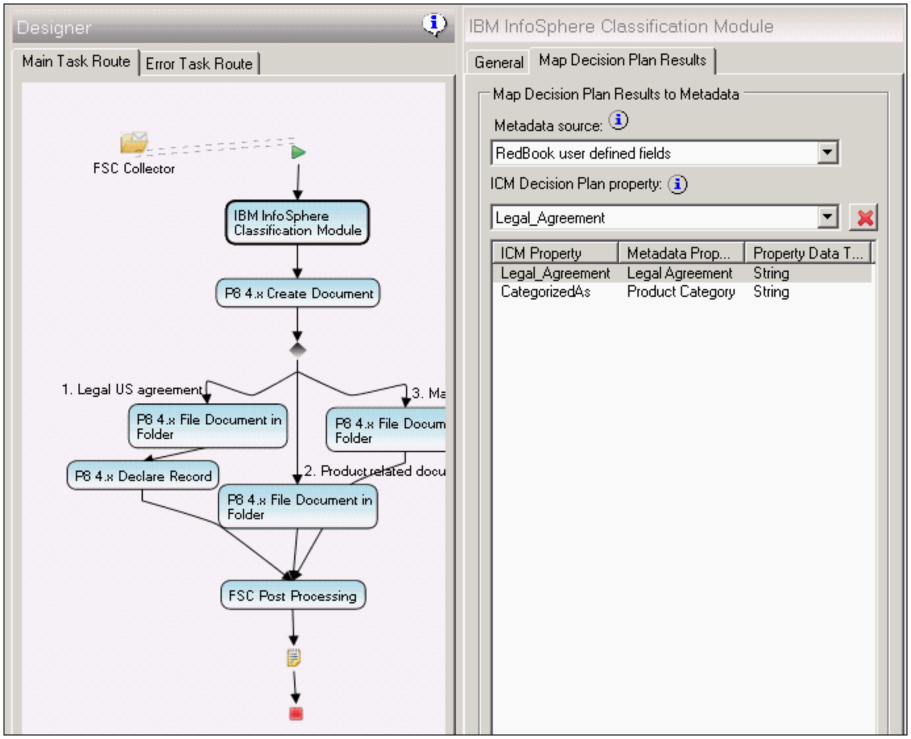


Figure 12-35 Completed task route

When invoking Classification Module, you must map the output fields from the decision plan to the Content Collector metadata fields, as shown in Figure 12-36.



Map Decision Plan Results to Metadata

Metadata source: RedBook user defined fields

ICM Decision Plan property: Legal_Agreement

ICM Property	Metadata Prop...	Property Data T...
Legal_Agreement	Legal Agreement	String
CategorizedAs	Product Category	String

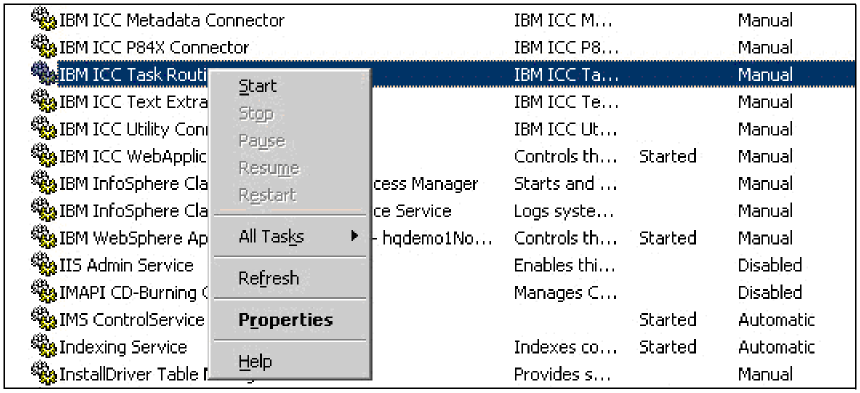
Figure 12-36 Mapping decision plan fields with Content Collector fields

The next step is to run Content Collector to collect the data in FileNet P8.

Collecting the exported Content Analytics data

The final step is to run the appropriate Content Collector services so that the task route that you just defined collects and stores the data exported by Content Analytics.

Open Services from the Windows toolbar, and start the **IBM ICC Task Routing Engine** service (Figure 12-37).



IBM ICC Metadata Connector	IBM ICC M...		Manual
IBM ICC P84X Connector	IBM ICC P8...		Manual
IBM ICC Task Routi	IBM ICC Ta...		Manual
IBM ICC Text Extra	IBM ICC Te...		Manual
IBM ICC Utility Conı	IBM ICC Ut...		Manual
IBM ICC WebApplic	Controls th...	Started	Manual
IBM InfoSphere Cla cess Manager	Starts and ...		Manual
IBM InfoSphere Cla ce Service	Logs syste...		Manual
IBM WebSphere Ap - hqdemo1No...	Controls th...	Started	Manual
IIS Admin Service	Enables thi...		Disabled
IMAPI CD-Burning (	Manages C...		Disabled
IMS ControlService		Started	Automatic
Indexing Service	Indexes co...	Started	Automatic
InstallDriver Table I	Provides s...		Manual

Figure 12-37 Running the Content Collector service

You can view the collected data in FileNet P8 that connects to the repository by using FileNet Workplace XT.

12.4 Records management and email archiving

An important part of a compliance ECM project is defining the Records Management framework, which includes the records classes and the file plan that governs them. In general, this project requires a good knowledge of regulatory rules, your company policies, and your content. You use IBM Content Collector integrated with Classification Module to inject the valuable business documents selected by the content assessment process into an ECM system and to declare records where necessary.

As part of the Content Assessment “preservation and organization” step, you define your file plan and the specific rules that govern the Records Declaration process of the content items. During the analysis step of the assessment process, you identify specific content groups (subsets) that you plan to organize in the ECM repository. At this stage, you use the two components of Content Assessment to proceed with the Records Management, Content Collector and Classification Module, as explained in the following steps:

1. Build a task route in Content Collector to implement the rules that you have defined for your Records Management.
2. Define a Classification Module decision plan with logic that relies on the content. As part of the Content Assessment phase, you have already

identified and exported a sample group of documents to illustrate the different categories of content:

   a. Export a sample set of documents for each category to a separate subfolder. Use this folder structure to build a Classification Module knowledge base that is to be used as part of the decision process.
   b. Build a Classification Module decision plan to incorporate the business logic that you decided to use to declare records. You define a series of meaningful fields to make the records declaration decision.

3. By using the Classification Module task, map the Classification Module result fields that contain the content-based decisions to the Content Collector metadata fields.
4. By using the Records Declaration task, execute the declaration of records at the document injection time.

Email archiving: For email archiving, you can use a similar workflow, assuming that you have Content Collector for Email. The Content Assessment focuses on crawling the email messages and analyzing them by using Content Analytics and Classification Module. After the assessment process is finalized, you can build the Classification Module decision plan, knowledge base, and the Content Collector task route in Content Collector for Email.

12.5 Preferred practices

Content Assessment is a complex task. Businesses are faced with a constant explosion of unstructured content. Companies are frequently running into barriers because of significant imposed user burdens. The text analysis function of Content Assessment gives you the tools to efficiently investigate the content and make informed decisions. The text miner application of Content Analytics (bundled with Content Assessment) plays a major role in this activity.

Consider the following tips to help you cope with this complex task:

- Use the Content Analytics component of Content Assessment to crawl the content sources. Take a small (if possible random) sample set of content in the initial exploration step so that you can take a quick view of the nature of your data. Follow the iterative process provided in 12.3, “Content Assessment workflow” on page 491.

- Inspect the data and try to determine the easy targets:
  - Understand the distribution of the content based on the file extension.
  - See if the file names have a pattern that can be used to further organize the content.
  - Activate Named Entity Recognition, and explore the facets it populates to understand if interesting business information can be inferred.
  - Activate *terms of interest* to identify relationships between nouns and nearby verbs and adverbs that exist in the content.
  - Activate *duplicate document detection* to identify documents that might contain the same or nearly the same content. The documents that have a duplicate document are good candidates for decommissioning and removing from your data set.
  - Engage *Document Clustering* to easily obtain immediate insight into your content.
- For initial inspection, use Content Analytics and the Classification Module annotator with the default knowledge bases and decision plans that are deployed, such as "Personal versus Business Content" or "Content Assessment."
- Inspect the exported sample data with the additional text miner tools provided by Classification Module. For example, use Taxonomy Proposer for Clustering with Classification Workbench to build your decision plan and train your knowledge bases according to the specific data discovered.

  Export a part of the data, and explore it in the Classification Module Workbench. As a result, build the Classification Module knowledge bases and decision plans by using Classification Module Workbench.
- In the Content Collector and Classification Module integration, use the new Classification Module knowledge base and decision plan that is tailored toward your compliance project business needs. The knowledge base and decision plan are for content that will be organized, enhanced with metadata, managed under Records Management according to the compliance policies of your company, and ready for eDiscovery.

12.6 Summary

Information chaos creates many challenges from expensive law suits to serious difficulties in managing your business efficiently. Content Assessment assists you in discovering valuable business information that is buried beneath irrelevant, obsolete, and duplicate content. It also helps you preserve and prepare your content for efficient eDiscovery.

13

Integrating Cognos Business Intelligence

IBM Cognos 8 Business Intelligence (BI) delivers a complete range of BI capabilities, including reporting, analysis, OLAP Cubes, dashboards, and scorecards. IBM Content Analytics adds value to the Cognos 8 BI product suite by making information available that is derived from your textual information by using its advanced text analytics. The information discovered by Content Analytics can be used to automatically generate predesigned Cognos reports from the text miner application. Alternatively, the data can be exported into a relational database. With this approach, you can design your own customized Cognos reports by using the Cognos Framework Manager and Cognos Advanced Report Studio.

This chapter begins with the initial steps for you to enable and configure connectivity between Content Analytics and Cognos 8 BI. Next it guides you through the process to generate predesigned Cognos reports from the text miner application. Finally, it outlines how to export Content Analytics data into a relational database so that Cognos can use the data and its associated model to build customized reports.

This chapter includes the following sections:

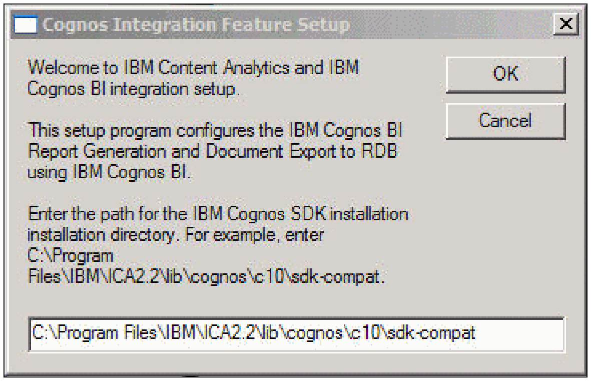
- Initial setup
- Generating Cognos BI reports
- Creating custom Cognos 8 BI reports

13.1 Initial setup

To integrate seamlessly with Cognos 8 BI, you must enable Cognos BI and provide the parameters that are necessary for communication. The following sections provide information about the initial setup steps.

13.1.1 Running the esrepcog command

Content Analytics uses the IBM Cognos software development kit (SDK) to programmatically communicate with IBM Cognos 8 BI and as such needs to detect where the SDK is installed. This information is provided by running the **`esrepcog`** command in the *`install location`*`/bin` directory. Figure 13-1 on page 526 shows the prompt after you run the **`esrepcog`** command. You can specify the directory of your own installed version of the Cognos 8 BI SDK or use the Cognos SDK packaged with Content Analytics. Enter the full path of the SDK as shown in Figure 13-1 on page 526.



Welcome to IBM Content Analytics and IBM Cognos BI integration setup.

This setup program configures the IBM Cognos BI Report Generation and Document Export to RDB using IBM Cognos BI.

Enter the path for the IBM Cognos SDK installation installation directory. For example, enter C:\Program Files\IBM\ICA2.2\lib\cognos\c10\sdk-compat.

C:\Program Files\IBM\ICA2.2\lib\cognos\c10\sdk-compat

Figure 13-1 Prompt for Cognos SDK installation location

You only need to run the **`esrepcog`** command once or after the location of the SDK has changed. After the **`esrepcog`** command has completed, restart the Content Analytics administration console as follows:

1. Open a command prompt (Windows) or Terminal (UNIX).
2. Enter the following commands:

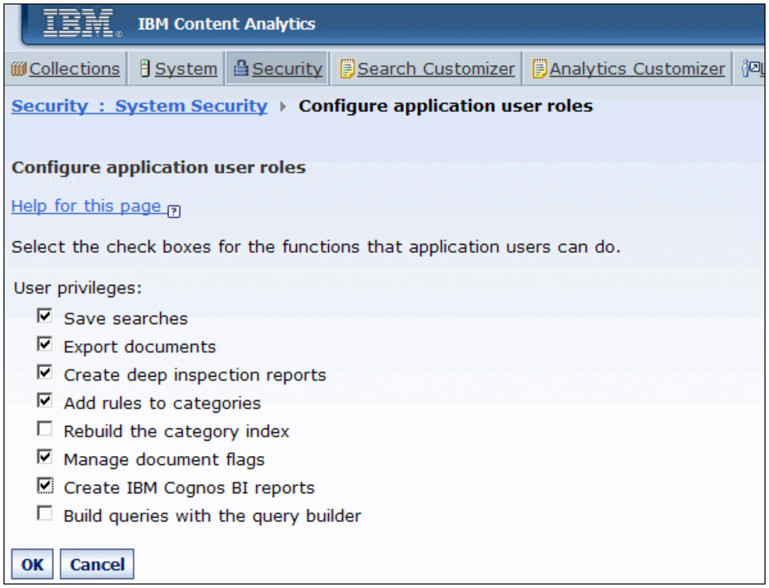
```
esadmin admin stop
esadmin admin start
```

The esrepcog command: By default, the IBM Cognos SDK for Cognos 8 BI packaged with Content Analytics is used for communication to Cognos. Therefore, if you plan to use Cognos 8 BI, you might not need to run the **`esrepcog`** command. If you plan to use another version of Cognos 8 BI, such as Cognos 10, run the **`esrepcog`** command, and specify the appropriate SDK. The SDK for Cognos 10 is packaged with Content Analytics and is in the `<install location>/lib/cognos/c10` directory.

13.1.2 Configuring default application user roles

The default application user role (*Analytics*) does not have privileges to create IBM Cognos BI reports. To assign privileges, follow these steps:

1. Select **System** → **System Security** → **Configure application user roles**.
2. In the Configure application user roles panel (Figure 13-2), under User privileges, select the following options and then click **OK**:
   - – Save searches
   - – Export documents
   - – Create deep inspection reports
   - – Add rules to categories
   - – Manage document flags
   - – Create IBM Cognos BI reports



Security : System Security ▸ **Configure application user roles**

Configure application user roles

Help for this page

Select the check boxes for the functions that application users can do.

User privileges:

☑ Save searches
☑ Export documents
☑ Create deep inspection reports
☑ Add rules to categories
□ Rebuild the category index
☑ Manage document flags
☑ Create IBM Cognos BI reports
□ Build queries with the query builder

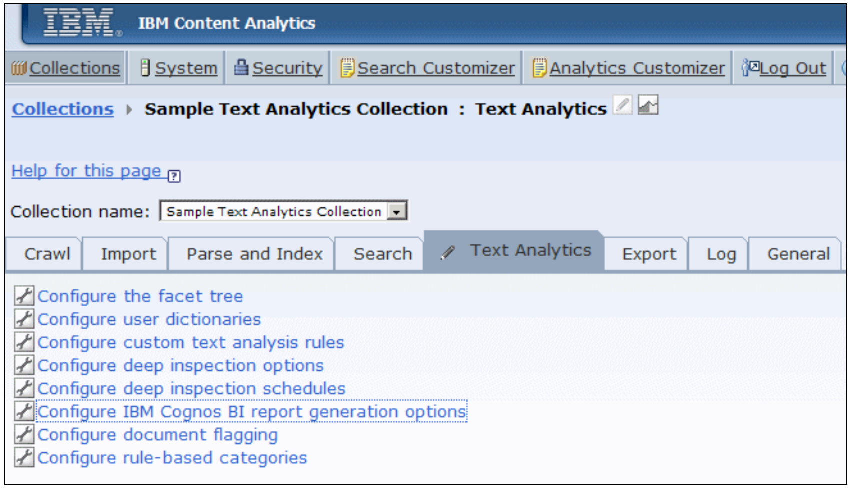
Figure 13-2 Enabling Cognos BI report generation for application

13.1.3 Configuring database connectivity

A relational database is used as the common store of information to be shared between Content Analytics and IBM Cognos 8 BI. This section guides you through the steps to inform Content Analytics about which relational database to use and how to connect to it.

To configure database connectivity, follow these steps:

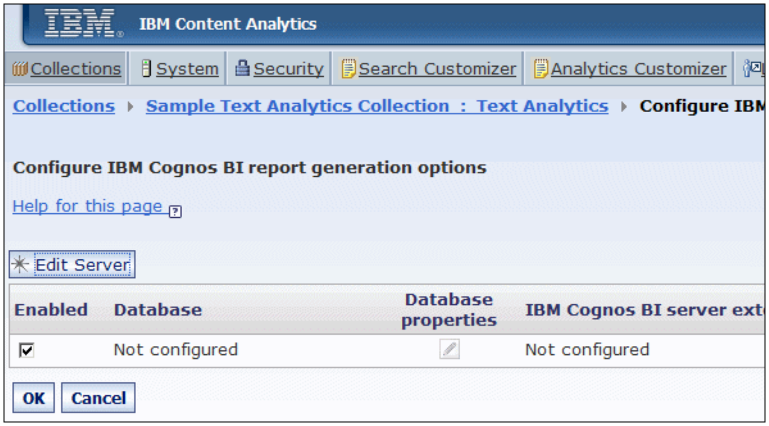
1. Select the **Text Analytics** tab (Figure 13-3).
2. Click **Configure IBM Cognos BI report generation options**.


IBM Content Analytics
Collections
System
Security
Search Customizer
Analytics Customizer
Log Out
Collections ▸ Sample Text Analytics Collection : Text Analytics
Help for this page
Crawl
Import
Parse and Index
Search
Text Analytics
Export
Log
General
Configure the facet tree
Configure user dictionaries
Configure custom text analysis rules
Configure deep inspection options
Configure deep inspection schedules
Configure IBM Cognos BI report generation options
Configure document flagging
Configure rule-based categories

Collection name: Sample Text Analytics Collection

Figure 13-3 Configure Cognos BI report generation option

3. In the Configure IBM Cognos BI report generation options panel (Figure 13-4), select the **Enabled** check box for the configuration entry, and click the **Edit Server** button.



Configure IBM Cognos BI report generation options

Enabled	Database	Database properties	IBM Cognos BI server ext
☑	Not configured		Not configured

Figure 13-4 Selecting the server to edit

4. In the Database Information panel (Figure 13-5 on page 530), complete these steps:

   a. Select the JDBC database type. In this example, we use **DB2**, which is installed locally on the same machine as Content Analytics.

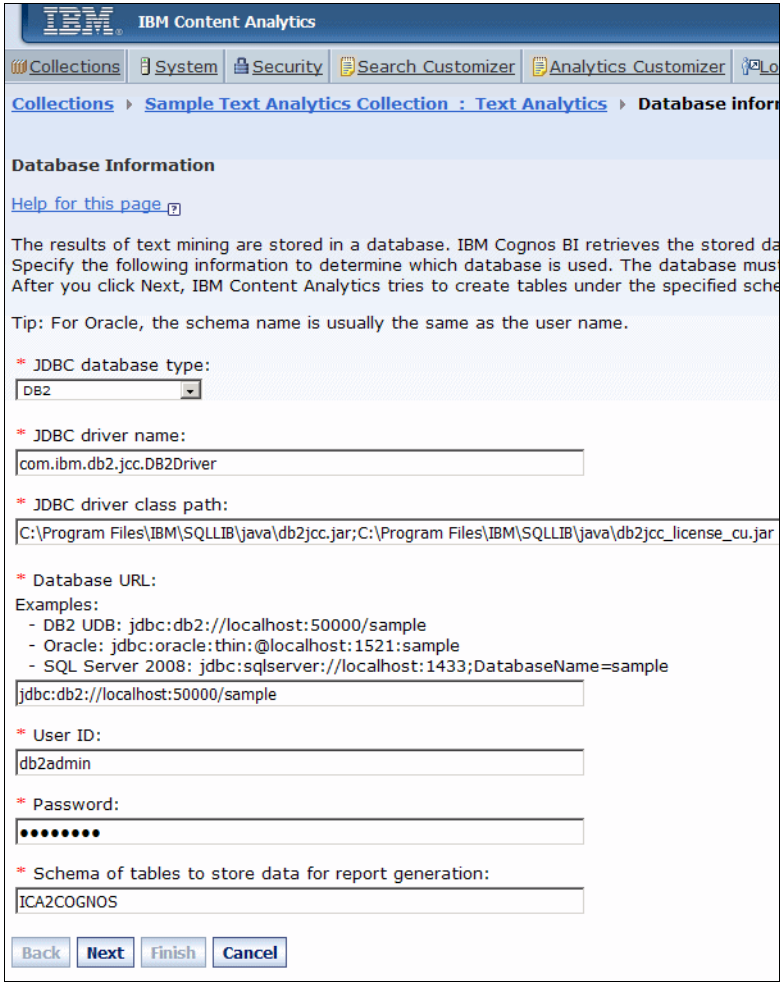
   b. Specify the JDBC driver name and class path if the values are different from the default values.

   c. Enter the database URL for your instance of the relational database. Examples are provide depending on the type of relational database chosen. The variable *localhost* is used in the examples and most likely must be changed to the host name of the database server.

   d. Provide a valid user ID and password.

   e. Enter a schema name of tables to store data for report generation. The schema name differentiates the Content Analytics/Cognos tables from other tables in the database (if any). This name can be anything you choose.

   f. Click **Next**.



Collections › Sample Text Analytics Collection : Text Analytics › Database infor

Database Information

Help for this page

The results of text mining are stored in a database. IBM Cognos BI retrieves the stored da
Specify the following information to determine which database is used. The database mus
After you click Next, IBM Content Analytics tries to create tables under the specified sche

Tip: For Oracle, the schema name is usually the same as the user name.

* JDBC database type:
DB2

* JDBC driver name:
com.ibm.db2.jcc.DB2Driver

* JDBC driver class path:
C:\Program Files\IBM\SQLLIB\java\db2jcc.jar;C:\Program Files\IBM\SQLLIB\java\db2jcc_license_cu.jar

* Database URL:
Examples:
- DB2 UDB: jdbc:db2://localhost:50000/sample
- Oracle: jdbc:oracle:thin:@localhost:1521:sample
- SQL Server 2008: jdbc:sqlserver://localhost:1433;DatabaseName=sample

jdbc:db2://localhost:50000/sample

* User ID:
db2admin

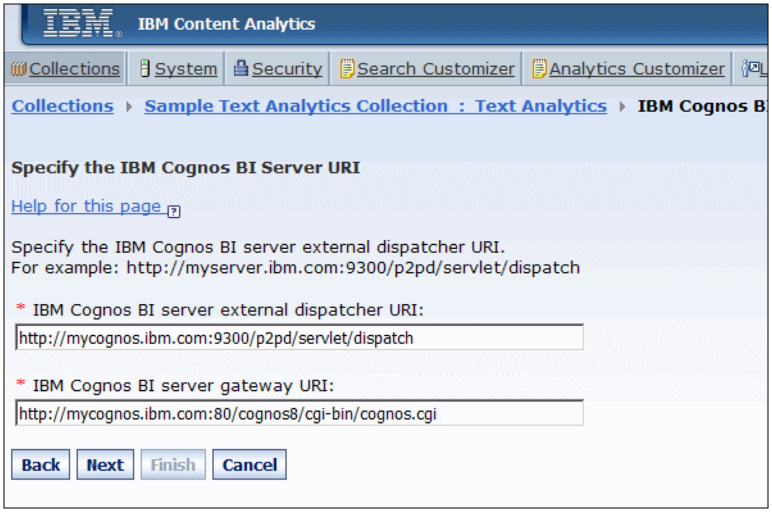
* Password:
••••••••

* Schema of tables to store data for report generation:
ICA2COGNOS

Figure 13-5 Providing database connectivity information

Content Analytics now attempts to connect to the specified relational database.

5. If Content Analytics connects to the relations database, in the Specify the IBM Cognos BI Server URI panel (Figure 13-6), specify the Uniform Resource Identifiers (URIs). Example URIs are provided by Content Analytics and most likely need to be changed. Obtain the correct URIs from your Cognos 8 BI Administrator. Click **Next**.



Specify the IBM Cognos BI Server URI

Specify the IBM Cognos BI server external dispatcher URI.
For example: http://myserver.ibm.com:9300/p2pd/servlet/dispatch

* IBM Cognos BI server external dispatcher URI: http://mycognos.ibm.com:9300/p2pd/servlet/dispatch

* IBM Cognos BI server gateway URI: http://mycognos.ibm.com:80/cognos8/cgi-bin/cognos.cgi

Figure 13-6 Specifying URIs to the Cognos 8 BI server

Content Analytics then connects to the Cognos BI Server and scans for the list of available data source connections.

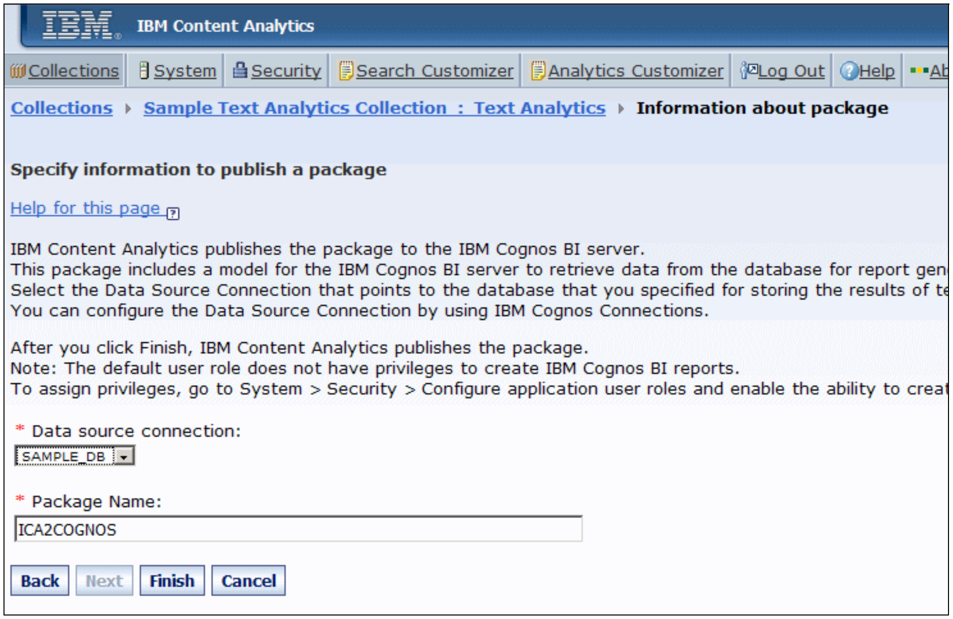
6. In the Specify Information to Publish a Package panel (Figure 13-7), complete these steps:

   a. Select the data source connection that points to the database that you just specified for storing the results of text mining. You can configure the data source connection in Cognos by using IBM Cognos Connections.

   b. Enter the package name to use.

   c. Click **Finish**.

   Content Analytics then publishes the package to the IBM Cognos BI server. This package includes a model for the IBM Cognos BI server to retrieve data from the database for report generation.



Specify information to publish a package

IBM Content Analytics publishes the package to the IBM Cognos BI server.
This package includes a model for the IBM Cognos BI server to retrieve data from the database for report gen
Select the Data Source Connection that points to the database that you specified for storing the results of te
You can configure the Data Source Connection by using IBM Cognos Connections.

After you click Finish, IBM Content Analytics publishes the package.
Note: The default user role does not have privileges to create IBM Cognos BI reports.
To assign privileges, go to System > Security > Configure application user roles and enable the ability to creat

* Data source connection:
SAMPLE_DB

* Package Name:
ICA2COGNOS

Figure 13-7 Specifying a package name

Privilege to create Cognos BI reports: The default application user role does not have privileges to create IBM Cognos BI reports. To assign privileges, go to **System** → **System Security** → **Configure application user roles** and enable the user privilege to create IBM Cognos BI reports.

You have now established connectivity between Content Analytics and Cognos 8 BI.

Restarting the Content Analytics server: After successfully creating a connection definition with the Cognos 8 BI server, you might need to stop and restart the Content Analytics server. By restarting the server, the Content Analytics search run time can properly establish a valid session Cognos 8 BI server.

13.2 Generating Cognos BI reports

Now that you have successfully set up the initial configuration for connectivity with Cognos 8 BI, you are ready to generate reports. This section explains how to generate the following predesigned Cognos 8 BI reports for each corresponding Content Analytics view in the text miner application:

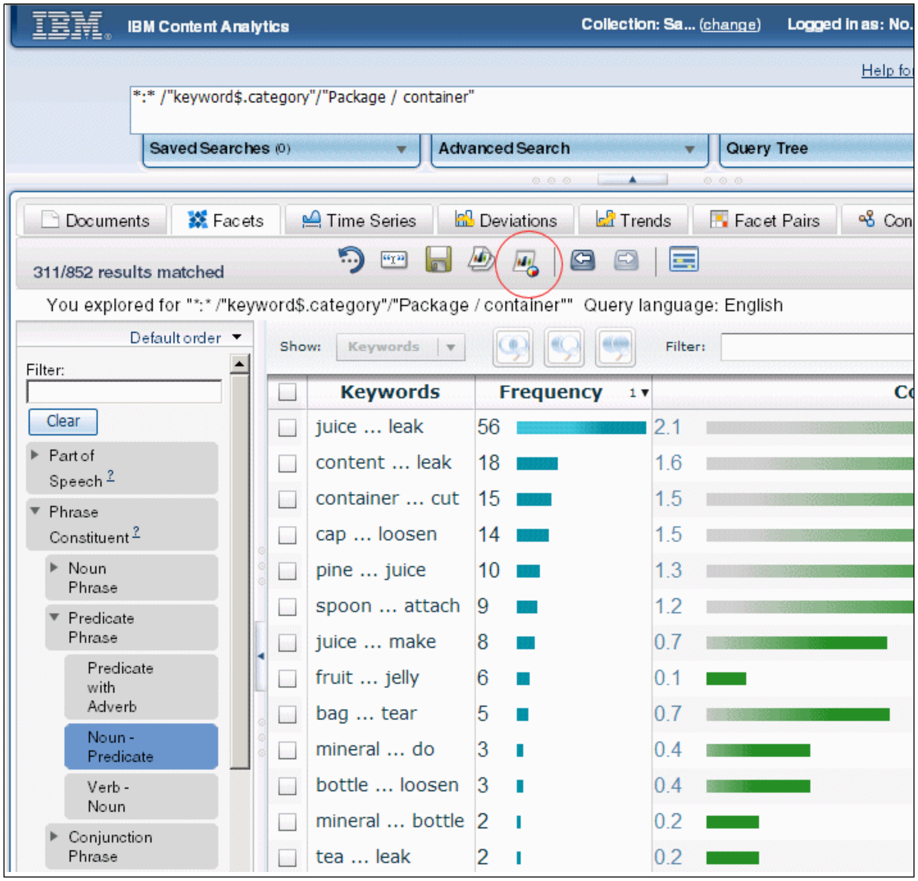
- Facets report
- Time series report
- Deviations report
- Trends report
- Facet pairs report

At anytime during the text miner discovery process, you can click a button to capture and generate a Cognos report that reflects the current state of your investigative work as seen through the current text miner view. For the sake of brevity, this section shows how to generate a Cognos 8 BI report from the facets view only. You use the same steps to generate reports from the other views. However, the result is a Cognos report that is designed for its corresponding text miner view.

Figure 13-8 on page 534 shows a facet view of noun predicate phrases in our sample text analytics collections that is restricted to the package-container category facet. (See the query expression at the top of the figure.) You might notice some interesting phrases indicating possible problems with the package containers. For example, a large number of containers indicate problems with leaking.

To generate a Cognos report that captures these potential problems, follow these steps:

1. Click the **Cognos report** icon in the lower-right corner of the menu bar of the search results (circled in Figure 13-8). As shown in the figure, the icon is displayed as a bar graph with a pie chart.



: /"keyword$.category"/"Package / container"

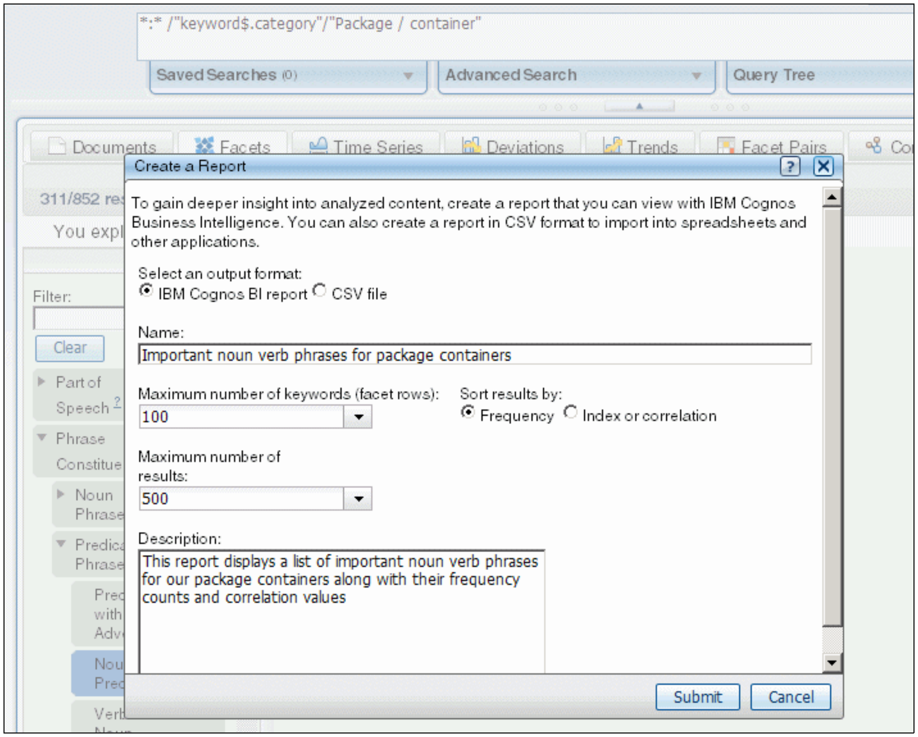
311/852 results matched

You explored for "*:* /"keyword$.category"/"Package / container"" Query language: English

Keywords	Frequency	Co
juice … leak	56	2.1
content … leak	18	1.6
container … cut	15	1.5
cap … loosen	14	1.5
pine … juice	10	1.3
spoon … attach	9	1.2
juice … make	8	0.7
fruit … jelly	6	0.1
bag … tear	5	0.7
mineral … do	3	0.4
bottle … loosen	3	0.4
mineral … bottle	2	0.2
tea … leak	2	0.2

Figure 13-8 Submitting a Cognos facet report request

2. In the Create a Report window (Figure 13-9), enter a meaningful name and optional description for your report. Verify that Select an output format is set to **IBM Cognos BI report** (the default setting). Then click **Submit**.



To gain deeper insight into analyzed content, create a report that you can view with IBM Cognos Business Intelligence. You can also create a report in CSV format to import into spreadsheets and other applications.

Select an output format:
● IBM Cognos BI report ○ CSV file

Name:
Important noun verb phrases for package containers

Maximum number of keywords (facet rows): 100

Sort results by: ● Frequency ○ Index or correlation

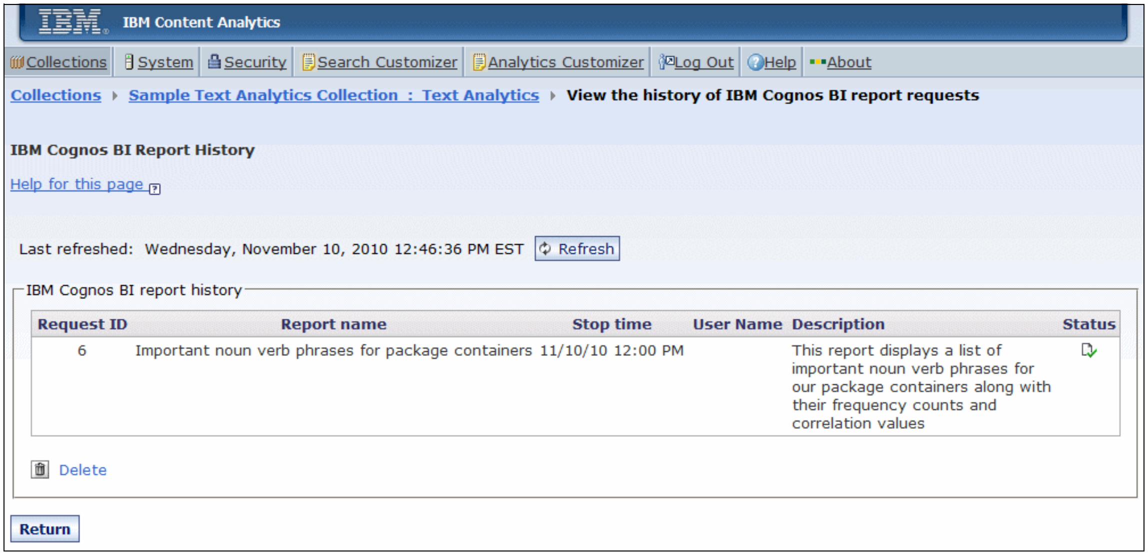
Maximum number of results: 500

Description:
This report displays a list of important noun verb phrases for our package containers along with their frequency counts and correlation values

Figure 13-9 Cognos report generation dialog box

A message box is displayed indicating that your report has been submitted for processing. The report is generated in the background. When it is ready, it is available for viewing from the **Report** tab of the text miner application.

As an administrator, you can view the status of all Cognos report submissions from the Content Analytics administration console under the **Text Analytics** tab while in monitor mode. Figure 13-10 shows the status of the facet view report of important noun-verb phrases for package containers. The green status icon on the far right side indicates a successful completion.


IBM Content Analytics
Collections
System
Security
Search Customizer
Analytics Customizer
Log Out
Help
About
Collections ▸ Sample Text Analytics Collection : Text Analytics ▸ View the history of IBM Cognos BI report requests

IBM Cognos BI Report History

Help for this page

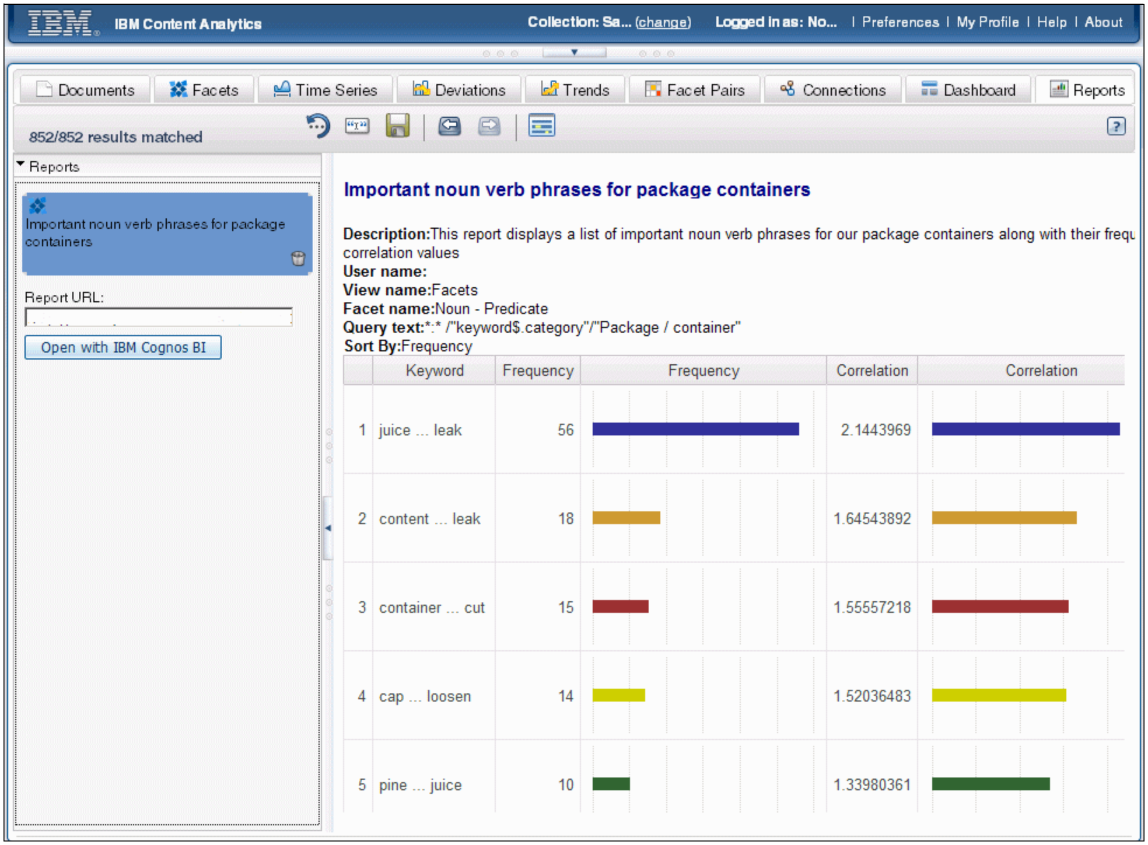
Last refreshed: Wednesday, November 10, 2010 12:46:36 PM EST Refresh

IBM Cognos BI report history

Request ID	Report name	Stop time	User Name	Description	Status
6	Important noun verb phrases for package containers	11/10/10 12:00 PM		This report displays a list of important noun verb phrases for our package containers along with their frequency counts and correlation values	✓

Figure 13-10 Checking the status of submitted Cognos report generation requests

For the business analyst working in the text miner application, the report normally takes about a minute or more to generate. When the report is completed, an entry for the report is displayed under the **Reports** tab in the text miner application. Figure 13-11 shows the facet report entry listed on the left side. When clicked, the Cognos generated report is displayed in the panel on the right side.



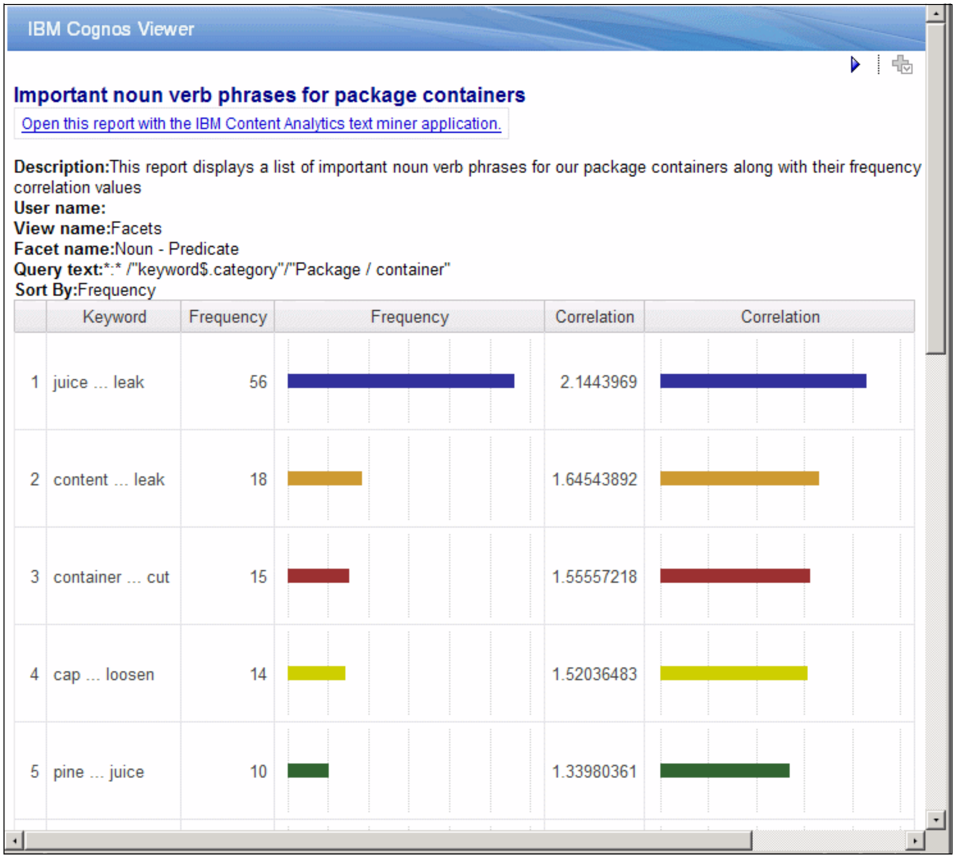
Important noun verb phrases for package containers

Description: This report displays a list of important noun verb phrases for our package containers along with their frequ correlation values
User name:
View name: Facets
Facet name: Noun - Predicate
Query text:*:* /"keyword$.category"/"Package / container"
Sort By: Frequency

	Keyword	Frequency	Frequency	Correlation	Correlation
1	juice ... leak	56	▬	2.1443969	▬
2	content ... leak	18	▬	1.64543892	▬
3	container ... cut	15	▬	1.55557218	▬
4	cap ... loosen	14	▬	1.52036483	▬
5	pine ... juice	10	▬	1.33980361	▬

Figure 13-11 Cognos report viewed through the text miner Reports tab

You can click the **Open with IBM Cognos BI** button on the left side to view the selected report entry in the native Cognos Report Viewer. Then a browser window opens with the Cognos Viewer displaying your reports as shown in Figure 13-12.



Important noun verb phrases for package containers

Open this report with the IBM Content Analytics text miner application.

Description: This report displays a list of important noun verb phrases for our package containers along with their frequency correlation values
User name:
View name: Facets
Facet name: Noun - Predicate
Query text:*:* /"keyword$.category"/"Package / container"
Sort By: Frequency

	Keyword	Frequency	Frequency	Correlation	Correlation
1	juice ... leak	56		2.1443969	
2	content ... leak	18		1.64543892	
3	container ... cut	15		1.55557218	
4	cap ... loosen	14		1.52036483	
5	pine ... juice	10		1.33980361	

Figure 13-12 Facet report as viewed in the Cognos Report Viewer

The Cognos report shown here includes a link underneath the report title. You can click this link to open the report in the Content Analytics text miner application. Then the text miner application is started with the same query that is used to generate the report, and the report is placed in the appropriate text miner view. With this approach, you can continue where you left off in your investigative work. By using the powerful features of these navigation techniques, you can conveniently jump back and forth between the two products.

13.3 Creating custom Cognos 8 BI reports

So far you have seen how Content Analytics works with Cognos 8 BI to automatically generate predesigned Cognos reports. The data for the reports is stored in the relational database that you identified during the initial setup with Cognos. The reports definitions are stored in the Cognos package that you also specified during the initial setup. Consequently, it is possible to access and modify these reports from the advanced business reporting modules of Cognos 8 BI.

A time might come when you want to build an entirely new Cognos report by using the text analytic information that was generated by Content Analytics. With Content Analytics support of this scenario, you can directly export content analytic data into a relational database and automatically generate an associated star schema for that data. With these two pieces in place, it is now business as usual for the Cognos report designer. The star schema describes the data and is used by the Cognos Framework Manager and report modules to design a broad array of reports. The data in the relational database is used to populate the newly designed reports.

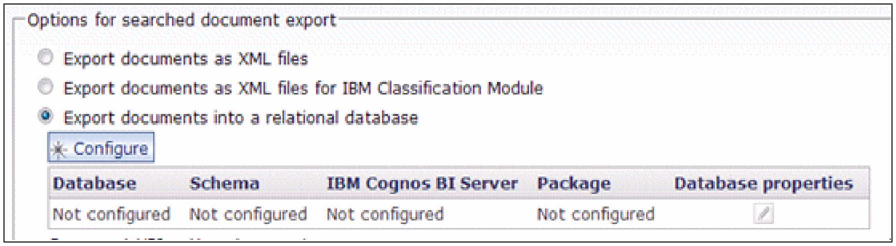
This section does not explain how to build customized Cognos reports. Instead it explains how to export data to a relational database from Content Analytics. It also explains where to find the corresponding star schema of the data. After these tasks are completed, a Cognos report designer can proceed with the creation of customized reports.

13.3.1 Configuring export options

To exporting data from Content Analytics, see Chapter 10, “Importing CSV files, exporting data, and performing deep inspection” on page 387. Content Analytics has three export points, namely after crawling, after indexing, and when exporting searched documents. For each of these export points, you can specify a relational database as the target of the exported results. When a relational database is chosen, the system further provides prompts about whether the exported results are intended for Cognos 8 BI.

The focus of this section is on the Cognos configuration of exported data. It begins by taking you through the configuration of the options for exporting searched documents, although the process is the same for the other two export points.

1. Click the **Export** tab for the collection to initiate the export configuration options.
2. Under Options for searched document export (Figure 13-13), select the appropriate option. In this example, we select the **Export documents into a relational database** option. Notice that the database, Cognos BI server, and package are not yet configured. Click the **Configure** button.



Options for searched document export

○ Export documents as XML files

○ Export documents as XML files for IBM Classification Module

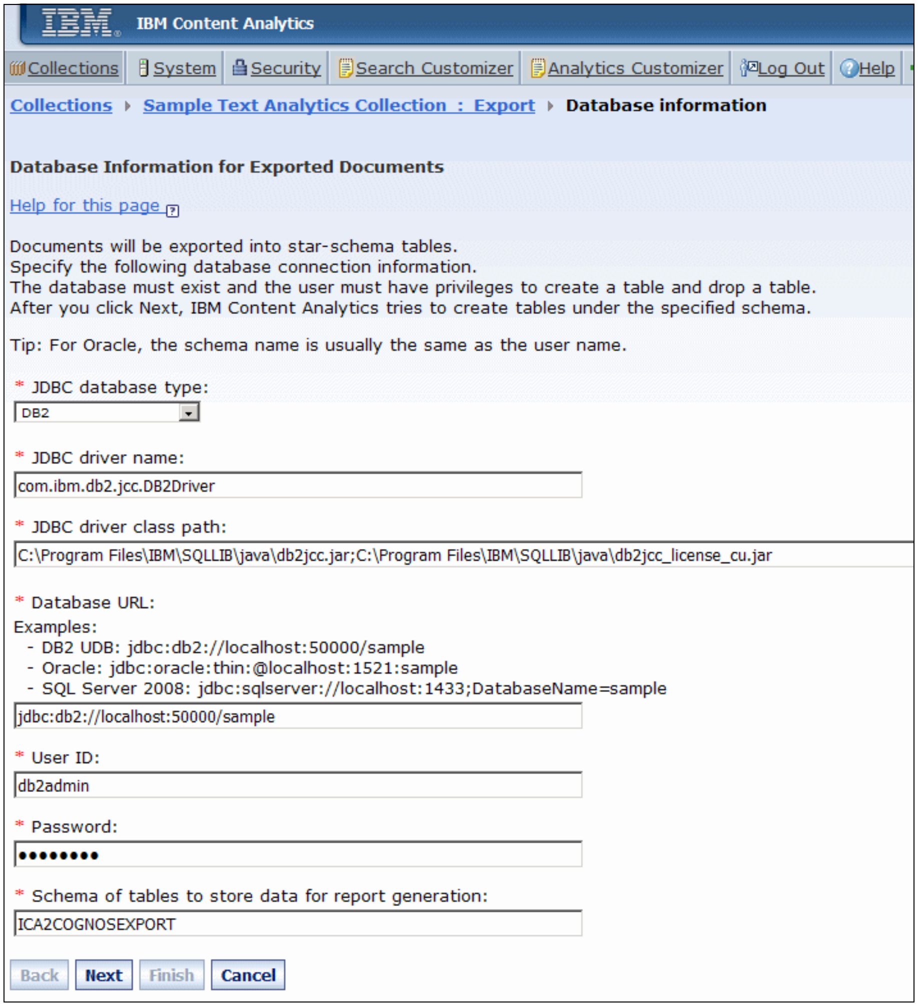
● Export documents into a relational database

Configure

Database	Schema	IBM Cognos BI Server	Package	Database properties
Not configured	Not configured	Not configured	Not configured	

Figure 13-13 Choosing to configure exported documents into a relational database

3. In the Database Information for Exported Documents panel (Figure 13-14 on page 541), complete the following steps:
   a. Select the JDBC relational database type. In this example, we select **DB2**, because it is installed locally on the same machine as Content Analytics.
   b. Enter the JDBC driver name and class path if they are different from the default values.
   c. Enter the database URL for your instance of the relational database. Several examples are provide depending on the type of relational database chosen. The variable *localhost* is used in the examples. You most likely need to change it to the host name of the database server.
   d. Enter a valid user ID and password.
   e. For Schema of tables to store data for report generation, enter a schema name for the Content Analytics and Cognos tables to be generated. The schema name differentiates the Content Analytics and Cognos tables from other tables in the database (if any).
   f. Click **Next**.



Collections ▸ Sample Text Analytics Collection : Export ▸ Database information

Database Information for Exported Documents

Help for this page

Documents will be exported into star-schema tables.
Specify the following database connection information.
The database must exist and the user must have privileges to create a table and drop a table.
After you click Next, IBM Content Analytics tries to create tables under the specified schema.

Tip: For Oracle, the schema name is usually the same as the user name.

* JDBC database type: DB2

* JDBC driver name: com.ibm.db2.jcc.DB2Driver

* JDBC driver class path: C:\Program Files\IBM\SQLLIB\java\db2jcc.jar;C:\Program Files\IBM\SQLLIB\java\db2jcc_license_cu.jar

* Database URL:
Examples:
- DB2 UDB: jdbc:db2://localhost:50000/sample
- Oracle: jdbc:oracle:thin:@localhost:1521:sample
- SQL Server 2008: jdbc:sqlserver://localhost:1433;DatabaseName=sample

jdbc:db2://localhost:50000/sample

* User ID: db2admin

* Password: ••••••••

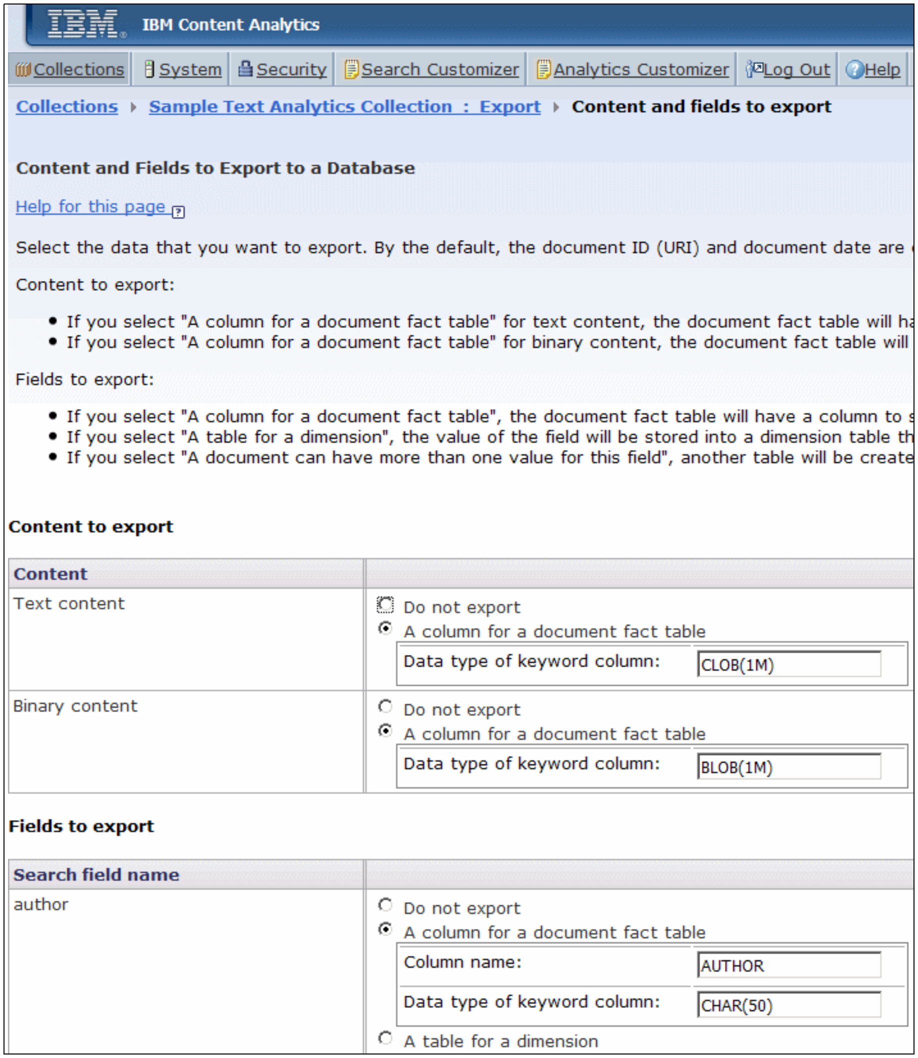
* Schema of tables to store data for report generation: ICA2COGNOSEXPORT

Figure 13-14 Configuring database connectivity for export option

Now Content Analytics attempts to connect to the specified relational database.

4. After Content Analytics connects to the specified relational database, in the Content and Fields to export to a Database panel (Figure 13-15), complete the fields for your collection to export.

   The first two fields represent the content of a document. The remaining fields represent the search fields configured for the collection. For each of these fields, select **A column for a document fact table** (in Cognos BI). After you select all of the fields that you want to configure for export, click **Next**.



Content and Fields to Export to a Database

Help for this page

Select the data that you want to export. By the default, the document ID (URI) and document date are

Content to export:

- If you select "A column for a document fact table" for text content, the document fact table will ha
- If you select "A column for a document fact table" for binary content, the document fact table will

Fields to export:

- If you select "A column for a document fact table", the document fact table will have a column to s
- If you select "A table for a dimension", the value of the field will be stored into a dimension table th
- If you select "A document can have more than one value for this field", another table will be create

Content to export

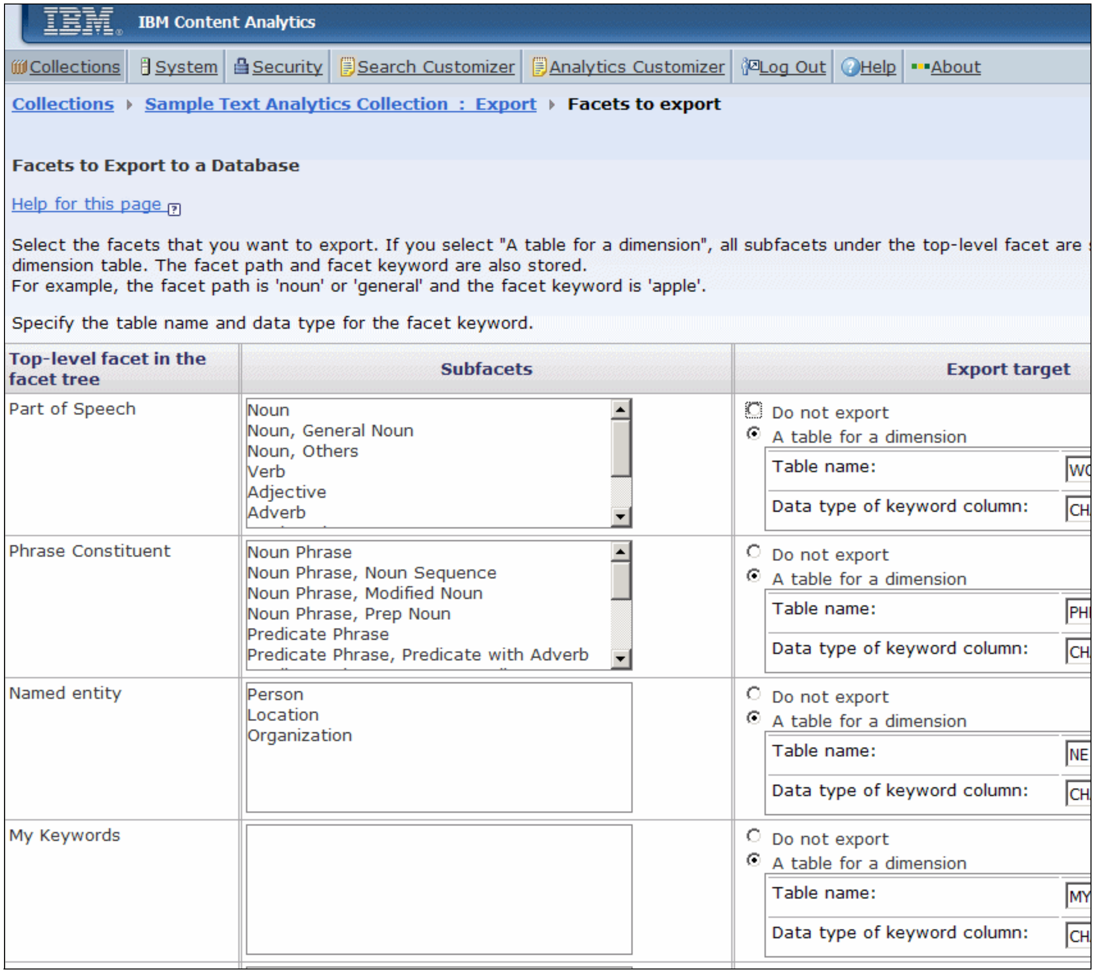
Content	
Text content	○ Do not export ● A column for a document fact table Data type of keyword column: CLOB(1M)
Binary content	○ Do not export ● A column for a document fact table Data type of keyword column: BLOB(1M)

Fields to export

Search field name	
author	○ Do not export ● A column for a document fact table Column name: AUTHOR Data type of keyword column: CHAR(50) ○ A table for a dimension

Figure 13-15 Configuring fields for export

5. In the Facets to Export to a Database panel (Figure 13-16), where you see the list of facets to be configured for export, select **A table for a dimension** (in Cognos BI) for each facet for export. Then, click **Next**.



Collections ▸ Sample Text Analytics Collection : Export ▸ Facets to export

Facets to Export to a Database

Help for this page

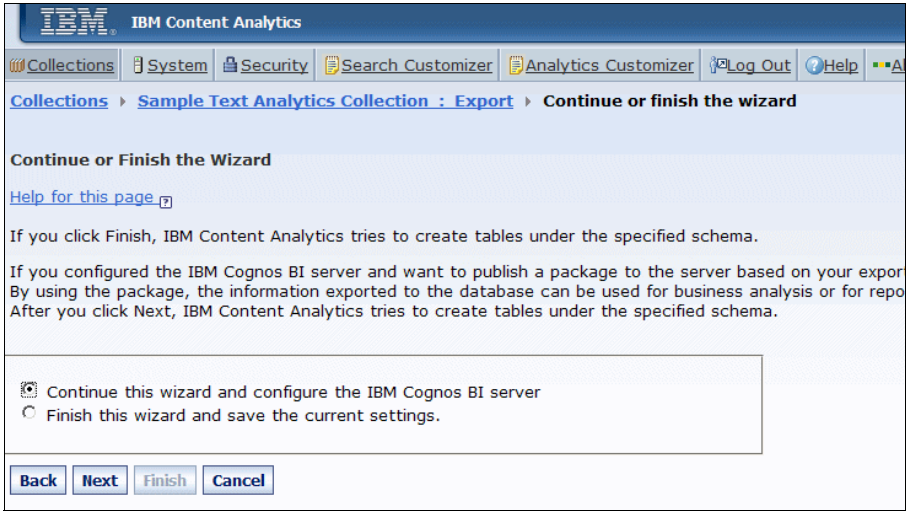
Select the facets that you want to export. If you select "A table for a dimension", all subfacets under the top-level facet are dimension table. The facet path and facet keyword are also stored.
For example, the facet path is 'noun' or 'general' and the facet keyword is 'apple'.

Specify the table name and data type for the facet keyword.

Top-level facet in the facet tree	Subfacets	Export target
Part of Speech	Noun Noun, General Noun Noun, Others Verb Adjective Adverb	○ Do not export ● A table for a dimension Table name: WC Data type of keyword column: CH
Phrase Constituent	Noun Phrase Noun Phrase, Noun Sequence Noun Phrase, Modified Noun Noun Phrase, Prep Noun Predicate Phrase Predicate Phrase, Predicate with Adverb	○ Do not export ● A table for a dimension Table name: PH Data type of keyword column: CH
Named entity	Person Location Organization	○ Do not export ● A table for a dimension Table name: NE Data type of keyword column: CH
My Keywords		○ Do not export ● A table for a dimension Table name: MY Data type of keyword column: CH

Figure 13-16 Configuring facets for export

6. In the Continue or Finish the Wizard panel (Figure 13-17), specify whether you want to continue configuration for the Cognos BI server or to finish the wizard and save the current settings. Accept the default setting **Continue with this wizard and configure the IBM Cognos BI server**. Click **Next**.



Continue or Finish the Wizard

Help for this page

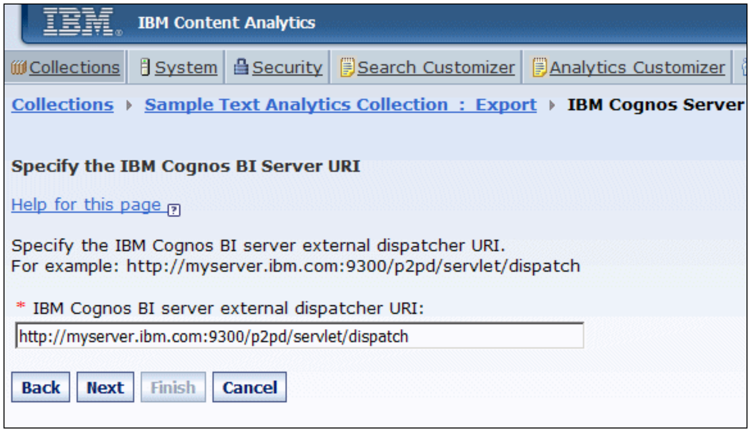
If you click Finish, IBM Content Analytics tries to create tables under the specified schema.

If you configured the IBM Cognos BI server and want to publish a package to the server based on your expor
By using the package, the information exported to the database can be used for business analysis or for repo
After you click Next, IBM Content Analytics tries to create tables under the specified schema.

● Continue this wizard and configure the IBM Cognos BI server
○ Finish this wizard and save the current settings.

Figure 13-17 Choosing to configure the Cognos BI server for export

7. In the Specify the IBM Cognos BI Server URI panel (Figure 13-18), verify whether the example URI is correct. You most likely must change it. You can obtain the correct URI from your Cognos 8 BI Administrator. Then click **Next**.



Specify the IBM Cognos BI Server URI

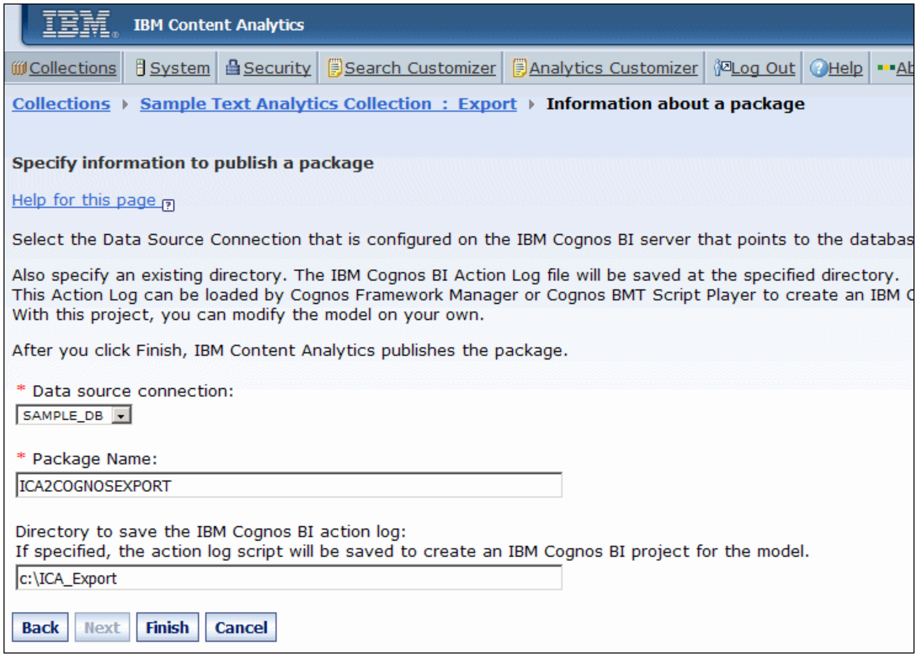
Help for this page

Specify the IBM Cognos BI server external dispatcher URI.
For example: http://myserver.ibm.com:9300/p2pd/servlet/dispatch

* IBM Cognos BI server external dispatcher URI:
http://myserver.ibm.com:9300/p2pd/servlet/dispatch

Figure 13-18 Prompt for Cognos BI server dispatcher URI

8. In the Specify information to publish a package panel (Figure 13-19), complete the following steps:
   a. Specify the data source connection to be used at the Cognos BI server. Make sure that the data source connection that you select is one that also connects to the same relational database that you specified as receiving the exported results.
   b. Enter a package name, which can be anything you choose and contains the model for your exported results in the Cognos BI server.
   c. Enter a directory to store the IBM Cognos BI action log script. The action log script can be used by the Cognos BI server to generate the star schema model for your data.
   d. Click **Finish**.


IBM Content Analytics
Collections | System | Security | Search Customizer | Analytics Customizer | Log Out | Help | Ab
Collections ▸ Sample Text Analytics Collection : Export ▸ Information about a package

Specify information to publish a package

Help for this page

Select the Data Source Connection that is configured on the IBM Cognos BI server that points to the databas

Also specify an existing directory. The IBM Cognos BI Action Log file will be saved at the specified directory.
This Action Log can be loaded by Cognos Framework Manager or Cognos BMT Script Player to create an IBM C
With this project, you can modify the model on your own.

After you click Finish, IBM Content Analytics publishes the package.

* Data source connection: SAMPLE_DB

* Package Name: ICA2COGNOSEXPORT

Directory to save the IBM Cognos BI action log:
If specified, the action log script will be saved to create an IBM Cognos BI project for the model.
c:\ICA_Export

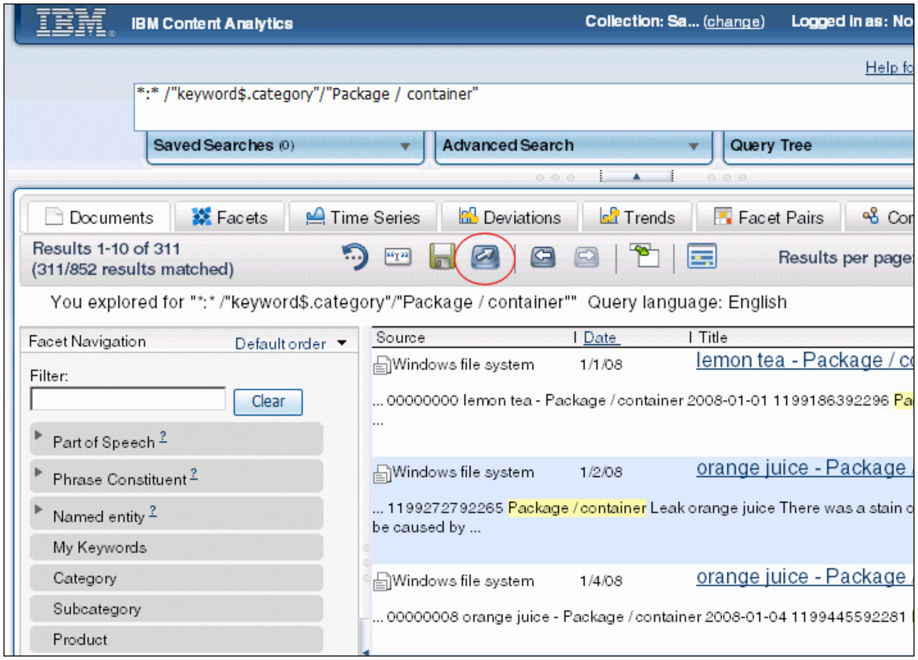
Figure 13-19 Cognos BI server configuration parameters for export

You have now successfully configured Content Analytics with the proper export parameters. The Cognos integration in this case has two major components. The first component is the relational database that is used as the repository of the exported content. The second component is the local directory where the corresponding activity log script will be deposited. Now you must export some documents.

13.3.2 Exporting search results

By way of example, this section explains how to export the same search result documents that are used in 13.2, “Generating Cognos BI reports” on page 533, but by using the **Export** button, not the **Report generation** button.

1. Click the **Export** button (highlighted in Figure 13-20).



: /"keyword$.category"/"Package / container"

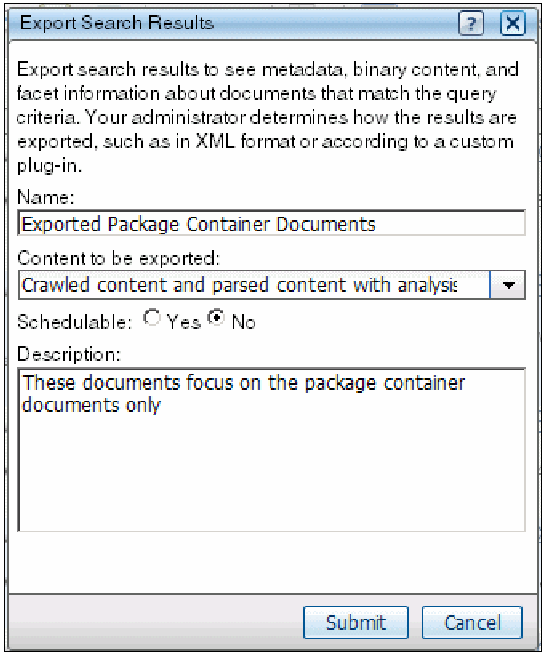
Results 1-10 of 311
(311/852 results matched)

You explored for "*:* /"keyword$.category"/"Package / container"" Query language: English

Source	Date	Title
Windows file system	1/1/08	lemon tea - Package / co
... 00000000 lemon tea - Package / container 2008-01-01 1199186392296 Pa ...		
Windows file system	1/2/08	orange juice - Package
... 1199272792265 Package / container Leak orange juice There was a stain o be caused by ...		
Windows file system	1/4/08	orange juice - Package
... 00000008 orange juice - Package / container 2008-01-04 1199445592281		

Figure 13-20 Exporting search result documents

2. In the Export Search Results dialog box (Figure 13-21), complete these steps:
   a. Enter a name for your export request.
   b. Specify the content to be exported. In this case, we are exporting the crawled content and the parsed and analysis data. All of the fields and facets specified during the previous export configuration are also exported.
   c. Optional: Enter a description.
   d. Click **Submit**.


Export Search Results

Export search results to see metadata, binary content, and facet information about documents that match the query criteria. Your administrator determines how the results are exported, such as in XML format or according to a custom plug-in.

Name:
Exported Package Container Documents

Content to be exported:
Crawled content and parsed content with analysis

Schedulable: ○ Yes ● No

Description:
These documents focus on the package container documents only

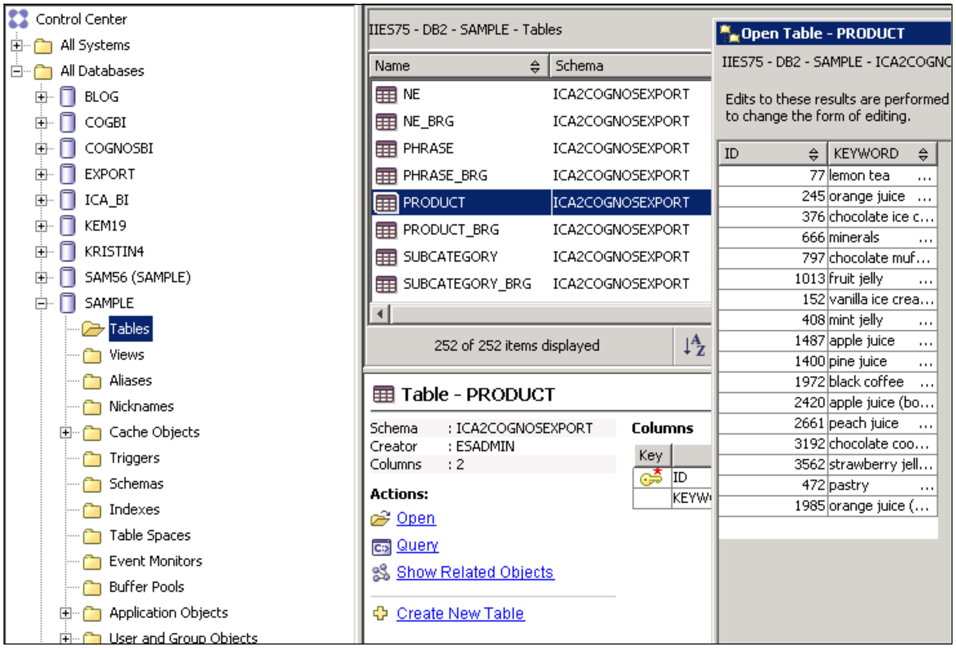
Figure 13-21 Search results export dialog

You then see a message indicating that your export request has been submitted for background processing. See 10.5, “Monitoring export requests” on page 407, to determine when your export request has been processed.

13.3.3 Loading the exported data model into Cognos

Your export request has now been successfully processed by Content Analytics. Now you must examine whether the export tables have been created and populated with data by using the DB2 Control Center. Figure 13-22 shows the product table in the sample database. As you can see on the right, the table is populated with a list of 17 product names.

Figure 13-22 also shows that many more tables were created by the export operation. Each table is assigned to the ICA2COGNOSEXPORT schema, which is the name provided for the schema during the configuration steps. Usage of this name is a convenient way to keep the export tables separate from other tables in the sample database.



Control Center
- All Systems
- All Databases
  - BLOG
  - COGBI
  - COGNOSBI
  - EXPORT
  - ICA_BI
  - KEM19
  - KRISTIN4
  - SAM56 (SAMPLE)
  - SAMPLE
    - Tables
    - Views
    - Aliases
    - Nicknames
    - Cache Objects
    - Triggers
    - Schemas
    - Indexes
    - Table Spaces
    - Event Monitors
    - Buffer Pools
    - Application Objects
    - User and Group Objects

IIES75 - DB2 - SAMPLE - Tables

Name	Schema
NE	ICA2COGNOSEXPORT
NE_BRG	ICA2COGNOSEXPORT
PHRASE	ICA2COGNOSEXPORT
PHRASE_BRG	ICA2COGNOSEXPORT
PRODUCT	ICA2COGNOSEXPORT
PRODUCT_BRG	ICA2COGNOSEXPORT
SUBCATEGORY	ICA2COGNOSEXPORT
SUBCATEGORY_BRG	ICA2COGNOSEXPORT

252 of 252 items displayed

Table - PRODUCT

Schema : ICA2COGNOSEXPORT
Creator : ESADMIN
Columns : 2

Columns

Key	
🔑	ID
	KEYW

Actions:
- Open
- Query
- Show Related Objects
- Create New Table

Open Table - PRODUCT

IIES75 - DB2 - SAMPLE - ICA2COGNC

Edits to these results are performed to change the form of editing.

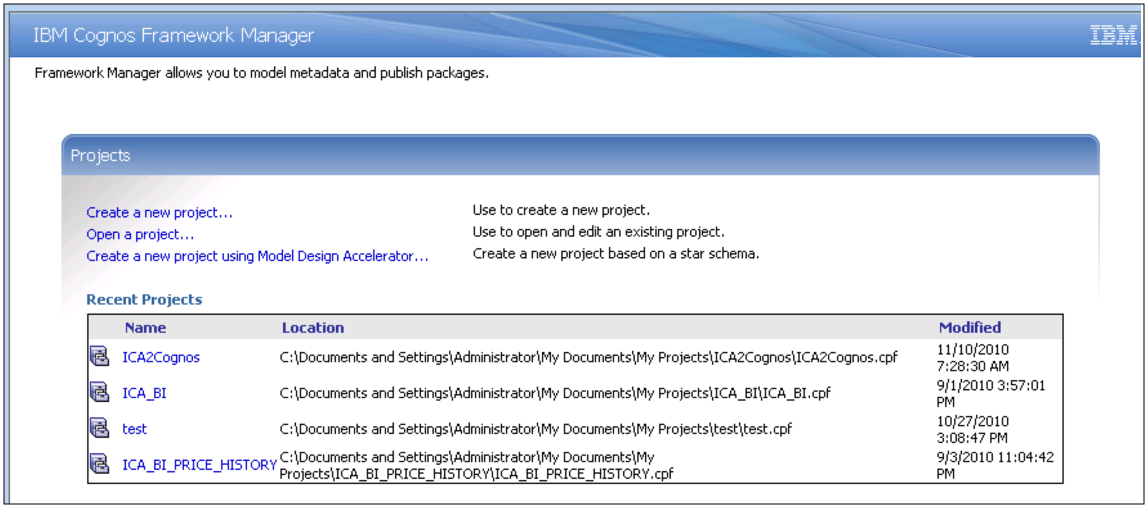
ID	KEYWORD
77	lemon tea ...
245	orange juice ...
376	chocolate ice c...
666	minerals ...
797	chocolate muf...
1013	fruit jelly ...
152	vanilla ice crea...
408	mint jelly ...
1487	apple juice ...
1400	pine juice ...
1972	black coffee ...
2420	apple juice (bo...
2661	peach juice ...
3192	chocolate coo...
3562	strawberry jell...
472	pastry ...
1985	orange juice (...

Figure 13-22 DB2 Control Center verification of export tables

You now have a relational database with several tables that are populated with data exported by Content Analytics. To generate custom Cognos reports on this data, you must first build a star schema by using the Cognos Framework Manager. The star schema describes all of the tables, their columns, and their relationships to Cognos. From this star schema, a Cognos report package can be published and made available to the Cognos Advanced and Business report modules.

To build a star schema, follow these steps:

1. Start Cognos Framework Manager (Figure 13-23).
2. Create a project. In this example, we create a project named *ICA2Cognos*.



IBM Cognos Framework Manager

Framework Manager allows you to model metadata and publish packages.

Projects

Create a new project... Use to create a new project.

Open a project... Use to open and edit an existing project.

Create a new project using Model Design Accelerator... Create a new project based on a star schema.

Recent Projects

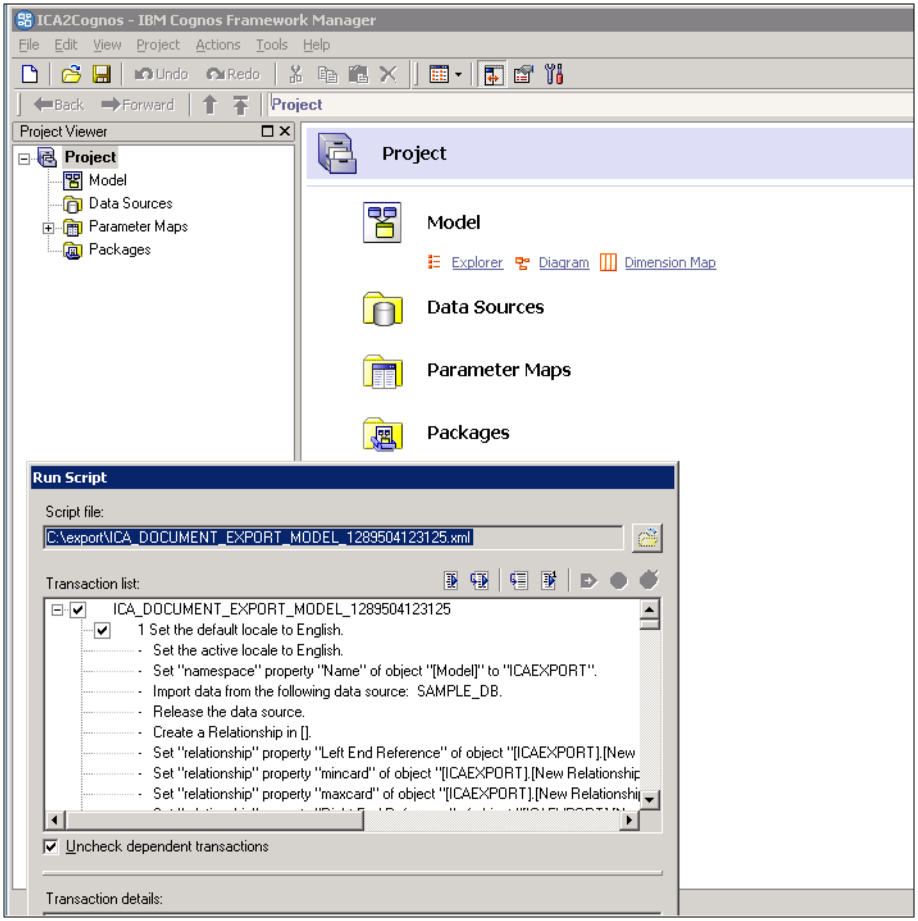
Name	Location	Modified
ICA2Cognos	C:\Documents and Settings\Administrator\My Documents\My Projects\ICA2Cognos\ICA2Cognos.cpf	11/10/2010 7:28:30 AM
ICA_BI	C:\Documents and Settings\Administrator\My Documents\My Projects\ICA_BI\ICA_BI.cpf	9/1/2010 3:57:01 PM
test	C:\Documents and Settings\Administrator\My Documents\My Projects\test\test.cpf	10/27/2010 3:08:47 PM
ICA_BI_PRICE_HISTORY	C:\Documents and Settings\Administrator\My Documents\My Projects\ICA_BI_PRICE_HISTORY\ICA_BI_PRICE_HISTORY.cpf	9/3/2010 11:04:42 PM

Figure 13-23 Cognos Framework Manager

3. Run the activity log script that was stored in the export directory that was specified during the configuration:

   a. Select **Project** → **Run Script**.

   b. In the Run Script window (Figure 13-24), select the XML file that was generated by Content Analytics. Click the **Accept** button to run the script.



Script file:

C:\export\ICA_DOCUMENT_EXPORT_MODEL_1289504123125.xml

Transaction list:

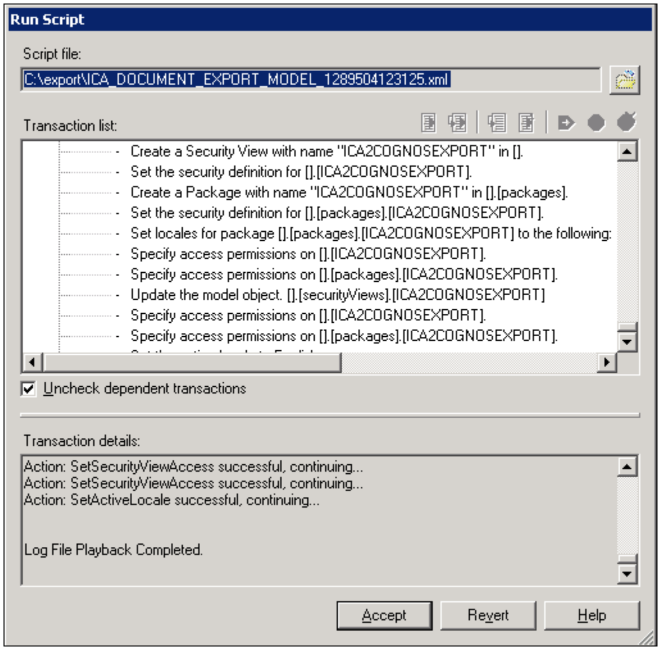
```
☑ ICA_DOCUMENT_EXPORT_MODEL_1289504123125
   ☑ 1 Set the default locale to English.
      - Set the active locale to English.
      - Set "namespace" property "Name" of object "[Model]" to "ICAEXPORT".
      - Import data from the following data source:  SAMPLE_DB.
      - Release the data source.
      - Create a Relationship in [].
      - Set "relationship" property "Left End Reference" of object "[ICAEXPORT].[New
      - Set "relationship" property "mincard" of object "[ICAEXPORT].[New Relationship
      - Set "relationship" property "maxcard" of object "[ICAEXPORT].[New Relationshi
```

☑ Uncheck dependent transactions

Transaction details:

Figure 13-24   Browsing for the activity log script file

Figure 13-25 shows the results of running the activity log script.



Script file: C:\export\ICA_DOCUMENT_EXPORT_MODEL_1289504123125.xml

Transaction list:

```
- Create a Security View with name "ICA2COGNOSEXPORT" in [].
- Set the security definition for [].[ICA2COGNOSEXPORT].
- Create a Package with name "ICA2COGNOSEXPORT" in [].[packages].
- Set the security definition for [].[packages].[ICA2COGNOSEXPORT].
- Set locales for package [].[packages].[ICA2COGNOSEXPORT] to the following:
- Specify access permissions on [].[ICA2COGNOSEXPORT].
- Specify access permissions on [].[packages].[ICA2COGNOSEXPORT].
- Update the model object. [].[securityViews].[ICA2COGNOSEXPORT]
- Specify access permissions on [].[ICA2COGNOSEXPORT].
- Specify access permissions on [].[packages].[ICA2COGNOSEXPORT].
```

☑ Uncheck dependent transactions

Transaction details:

```
Action: SetSecurityViewAccess successful, continuing...
Action: SetSecurityViewAccess successful, continuing...
Action: SetActiveLocale successful, continuing...


Log File Playback Completed.
```

Figure 13-25 Running the activity log script

4. Click the **Diagram** tab as shown in Figure 13-26 to view the generated star schema.

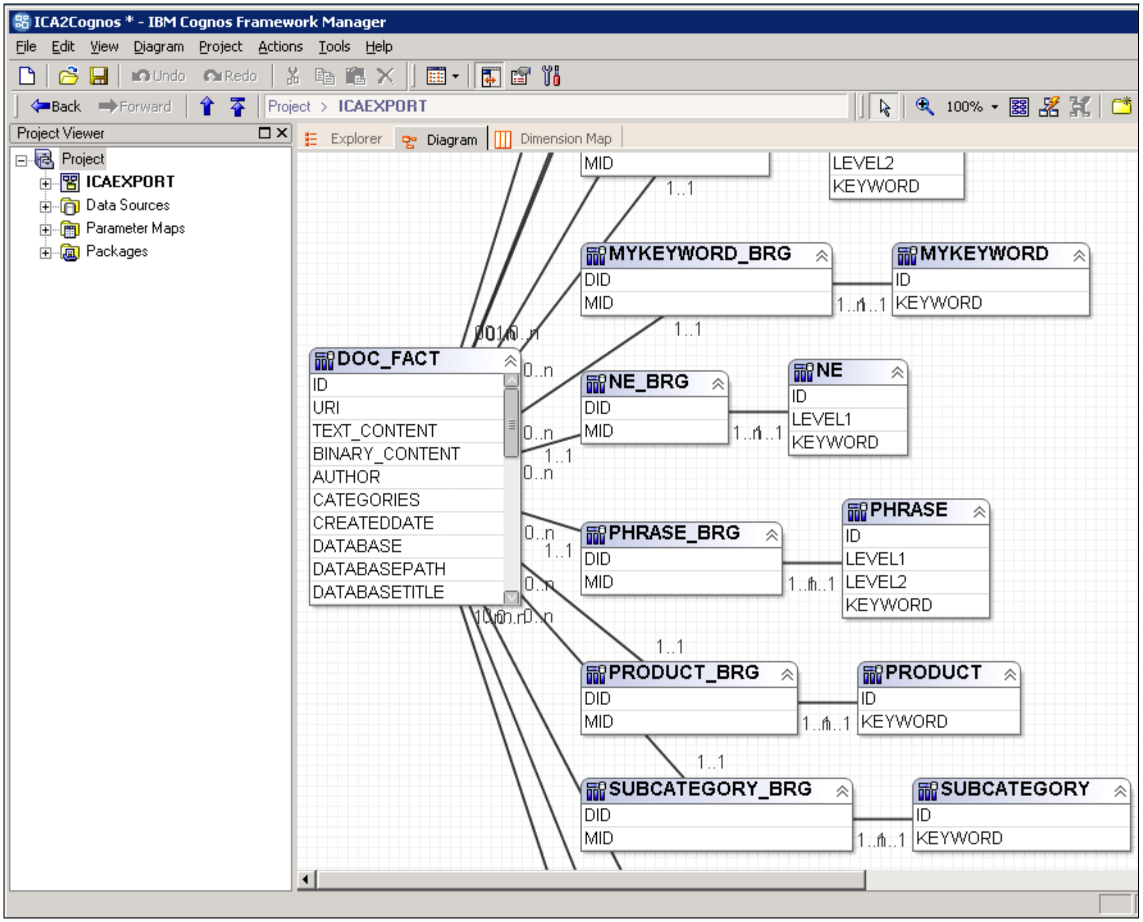
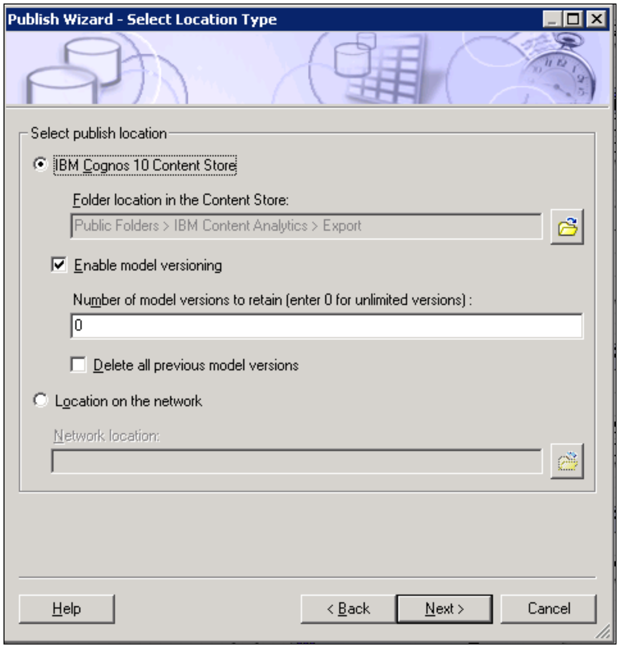


Figure 13-26 Star schema generated for Content Analytics exported data

5. Publish the package for this star schema. Select **Actions** → **Package** → **Publish packages**.

6. In the Publish Wizard - Select Location Type window (Figure 13-27), select **IBM Cognos 8 Content Store**, enter a folder location, and then click **Next**.



Publish Wizard - Select Location Type

Select publish location

● IBM Cognos 10 Content Store

Folder location in the Content Store: Public Folders > IBM Content Analytics > Export

☑ Enable model versioning

Number of model versions to retain (enter 0 for unlimited versions) : 0

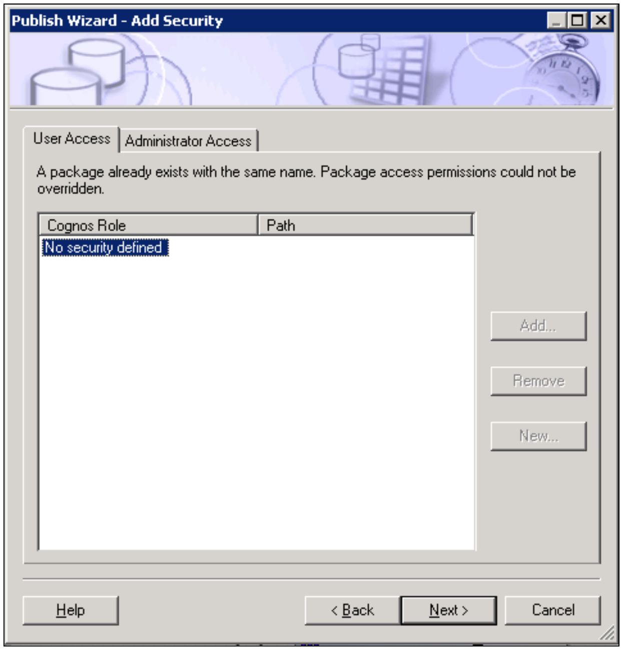
□ Delete all previous model versions

○ Location on the network

Network location:

Figure 13-27 Publishing your package to a location

7. In the Publish Wizard - Add Security window (Figure 13-28), accept the defaults for security and click **Next**.



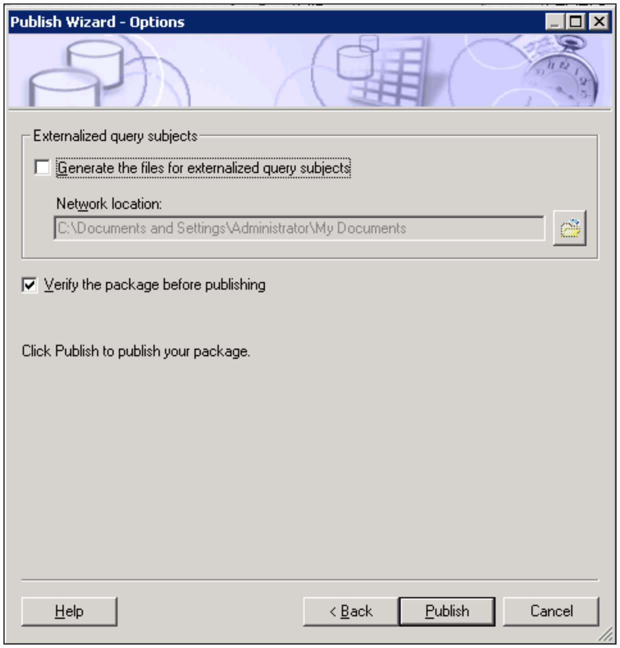
Publish Wizard - Add Security

A package already exists with the same name. Package access permissions could not be overridden.

Cognos Role	Path
No security defined	

Figure 13-28   Setting package access rights

8. In the Publish Wizard - Options window (Figure 13-29), accept the default to verify the package, and then click **Publish**.


Publish Wizard - Options

Externalized query subjects

□ Generate the files for externalized query subjects

Network location: C:\Documents and Settings\Administrator\My Documents

☑ Verify the package before publishing

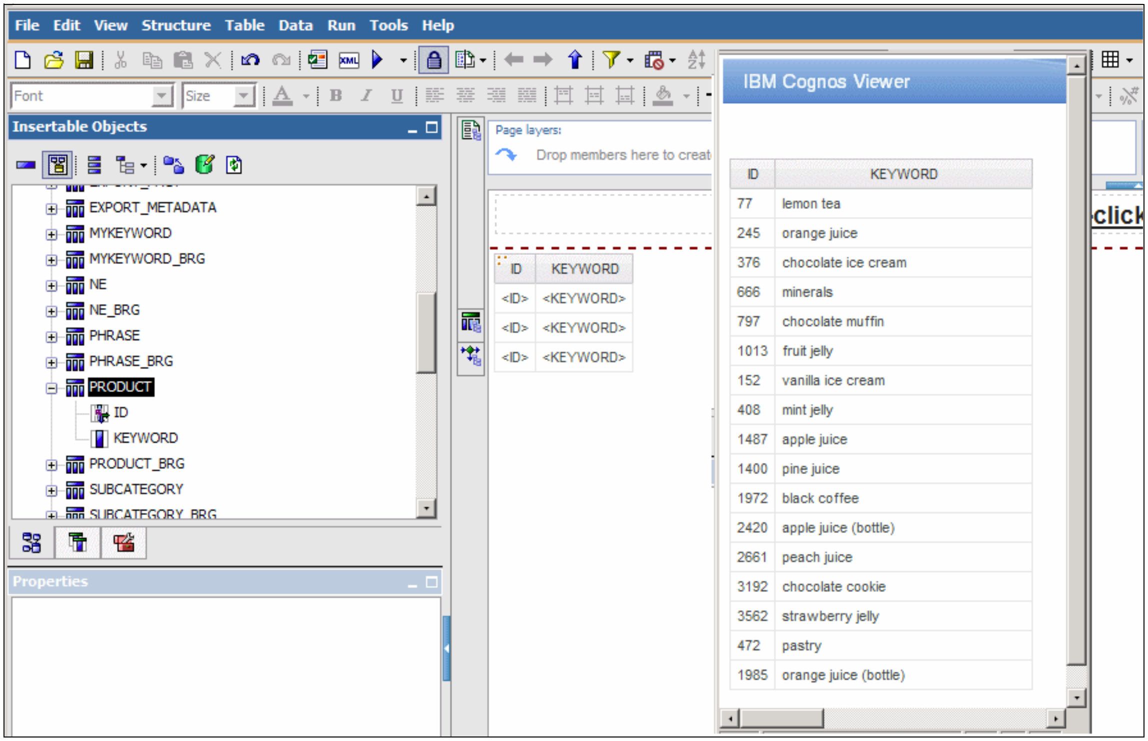
Click Publish to publish your package.

Figure 13-29 Verify the package before publishing

Now that the ICA2COGNOSEXPORT package has been published, you can build a custom report in the Cognos Report Studio:

1. Load the report package, which is the ICA2COGNOSEXPORT report package in this example.
2. Select a simple list report to be created.
3. Drag the product table to the list report template.

4. Select **Run** → **Run report-HTML**. The list report of the product table is then displayed in the IBM Cognos Viewer as shown on the right side in Figure 13-30.



ID	KEYWORD
<ID>	<KEYWORD>
<ID>	<KEYWORD>
<ID>	<KEYWORD>

ID	KEYWORD
77	lemon tea
245	orange juice
376	chocolate ice cream
666	minerals
797	chocolate muffin
1013	fruit jelly
152	vanilla ice cream
408	mint jelly
1487	apple juice
1400	pine juice
1972	black coffee
2420	apple juice (bottle)
2661	peach juice
3192	chocolate cookie
3562	strawberry jelly
472	pastry
1985	orange juice (bottle)

Figure 13-30 List report of Content Analytics product data

14

Customizing and extending the text miner application

The text miner application is a powerful text analysis tool that provides invaluable insight to your unstructured content. The text miner application provides many different views from which to analyze your data, ranging from views for time series and trend pattern analysis to various facet correlation views. You can customize the text miner application or extend it. For example, you can add one or more of your own text analytic views that are customized to your specific data and visualization needs. This chapter addresses how to customize and extend the application.

This chapter includes the following sections:

- Customizing the text miner application
- Reasons for extending the text miner application
- Sample plug-ins for text miner views
- Customizing the sample text miner plug-in
- Testing the customized plug-in

14.1 Customizing the text miner application

The previous chapters show you how to use the text miner application. This section provides tips to help you customize the text miner application to meet your business requirement. Customizing the text miner application is useful when you examine the text miner application during the testing stage. Do *not* modify the text miner application after the system goes into production.

> **Limiting user access to specific collections:** You can limit user access to specific collections for security purposes. To limit user access, see "Limiting user access to the text analytics collection" on page 647.

14.1.1 Analytics Customizer

Content Analytics provides the Analytics Customizer application to help you customize parts of the text miner application for your company. The advantage of using Analytics Customizer is that you can quickly update the properties that are used frequently by using a GUI. You do not have to edit the configuration file directly. You can examine and change the properties during the testing period.

> **Setting the customizerDisabled property:** Set the customizerDisabled property to `true` after you finalize the customization of your text miner application and prepare the system to go to production.

Accessing the Analytics Customizer depends on your deployment. If you use the deployed Jetty web server, click the **Analytics Customizer** link in the administration console (Figure 14-1).

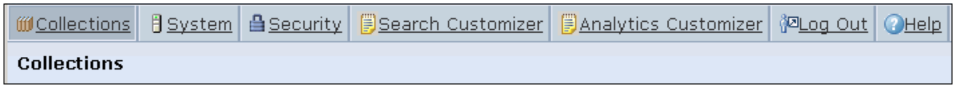


Figure 14-1 Linking to the Analytics Customizer in the administration console

When you open the Text Analytics Customizer, you can update the following options on each tab:

Server Specify the host name, protocol (HTTP or HTTPS), logging level, and query timeout in seconds.

Screen Specify images and texts, links, and paths to view in the window.

Query options Specify the **Query Options** tab or the **File Type filterResults** tab.

Results Specify what to show in the search result.

Images Specify the image files for the data sources.

You can also set the Preferences for the views.

After you update the values, click **Close** in the Analytics Customizer window, and then click **Exit**.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *customizing applications*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

14.1.2 Modifying the URI link in the Documents view

In the Documents view, the text miner application shows the document link that you can click to view the data in the data source. The document link is constructed based on the indexed Uniform Resource Identifier (URI).

Depending on the data source, the URI stored in the index is different as described in the “URI formats in the index” topic in the IBM Content Analytics Information Center at the following address:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Sometimes you need to change the document link so that it does not direct you to the data source. Alternatively, you might want to remove the document link from the search result in the Documents view. In this case, the regular expression URL filter helps to achieve your requirement. If you configure the regular expression URL filter properly, you can replace the URL that is used in the document link.

For example, you store the original data as an XML file in the file system. The XML file is stored as the following URI in the index:

```
file://c:/shared/document1.xml
```

To show the file from a web server so that it is easier to read and the system does not have to fetch the XML file directly from the document link, use the following link:

```
http://example.com/data/document1.xml
```

Displaying the XML tasks: If you want to use this example, you must configure the web server properly to display the XML files in a browser beforehand. This task is independent.

In this example, you can set the regular expression filter as follows:

```
|^file://c:/shared|http://example.com/data|
```

You can disable all document links so that users cannot click to see the data source. To disable all document links, set the regular express filter as `|.*||`, which means that you must replace any string (`.*`) with `null`.

After you determine how you want to change the URL that is used in the document link, you must perform additional tasks in the Content Analytics configuration. For more information, go to the IBM Content Analytics Information Center at the following address, and search on *configuring a regular expression URL filter for search results*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

14.2 Reasons for extending the text miner application

You might extend the text miner application for one of two reasons.

First, you might use several tools (of which Content Analytics is only a part) in the analysis of your data. If these tools are accessible by using a web browser or provide a programmable GUI, you can incorporate these tools into the text miner application and make them accessible from their own tabbed views. In this way, the text miner application consolidates most, if not all, of your various analytic tools into a single portal. This approach greatly simplifies the switching back and forth between tasks.

The second reason is that you realize that a more suitable visualization technique might better serve your data and that you have the programmatic means to implement such a view. In this case, you want to create your own view of the data provided by Content Analytics. To access the data, you can use the Content Analytics REST API to search and retrieve any of the information stored in a text analytic collection. When the information retrieved, you might have your own programmatic means to manipulate and display the data.

The remainder of this chapter concentrates on the latter use case and demonstrates an example extension of the text miner application.

14.3 Sample plug-ins for text miner views

Extending the text miner application follows a plug-in architecture where each plug-in creates a view tab on the results menu bar. The analyticsViewPlugin samples in the `ES_INSTALL_ROOT/samples/` directory provide the following example plug-ins:

myFirstPlugin A simple plug-in that uses the Dojo Toolkit only. The Dojo toolkit is already installed with the text miner web application.

mySecondPlugin A simple plug-in that uses the Dojo Toolkit and Adobe® Flex files.

TiaraPlugin An example of an interactive, visual text summarization view that shows the topic keywords for facets over a specified time period.

In this chapter, you modify the first sample plug-in by adding your own visualizations of the data. Therefore, you must perform the following steps to enable the first sample plug-in in the text miner application:

1. Copy the entire contents of the `samples/analyticsViewPlugin` directory into the `ES_NODE_ROOT/master_config/searchapp/analytics/plugin` directory. If the `plugin` directory does not exist, create one.
2. In the `plugins.xml` configuration file, uncomment the definitions for the myFirstPlugin sample.
3. Restart the text miner application:
   - If you use the provided Jetty web server, enter the following command, where *node_ID* identifies the search server:

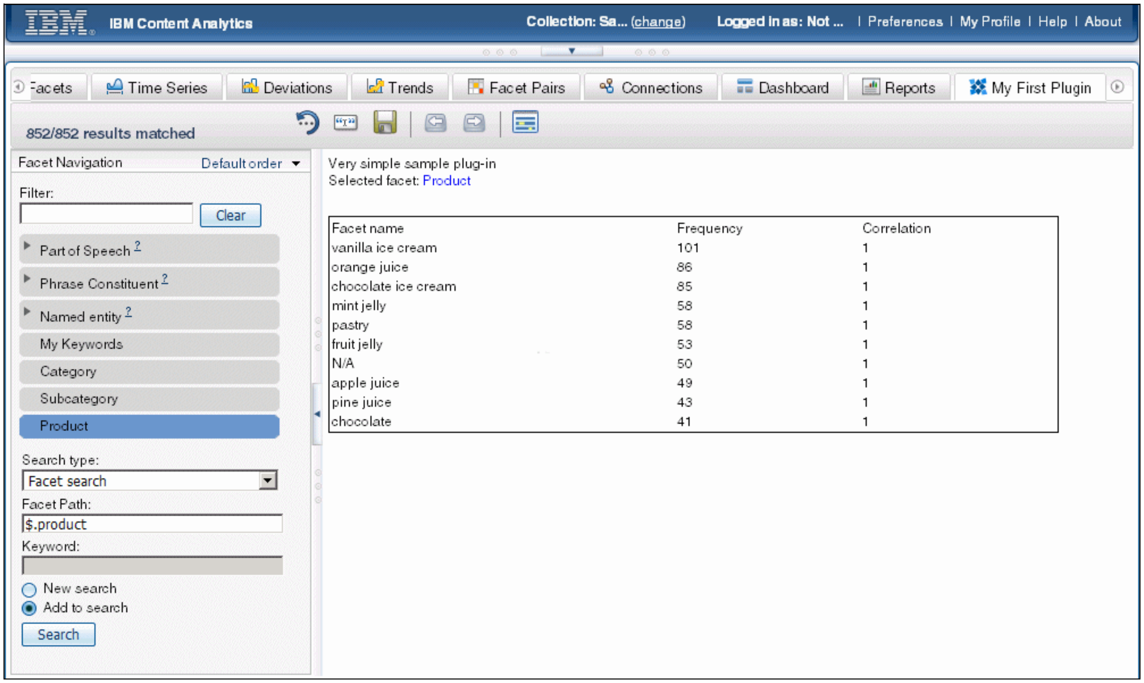
     ```
     esadmin session searchapp.node_ID restart
     ```

   - If you use WebSphere Application Server, enter the following command:

     ```
     esadmin config sync
     ```

4. Stop and restart the text miner application.

The myFirstPlugin sample is displayed to the right as the last text mining view that is provided with the product (Figure 14-2). (We use the sample text analytics collection throughout this example demonstration.) To the right of the **Reports** tab is the highlighted **My First Plugin** tab. In the Facet Navigation pane, the Product facet is the currently selected facet. A table is displayed on the right side that shows the list of its facet values, their document frequency counts, and associated correlation values.



852/852 results matched

Filter:

Search type: Facet search

Facet Path: $.product

Keyword:

○ New search
● Add to search

Very simple sample plug-in
Selected facet: Product

Facet name	Frequency	Correlation
vanilla ice cream	101	1
orange juice	86	1
chocolate ice cream	85	1
mint jelly	58	1
pastry	58	1
fruit jelly	53	1
N/A	50	1
apple juice	49	1
pine juice	43	1
chocolate	41	1

Figure 14-2 MyFirstPlugin enabled in the text miner application

You have now successfully enabled your first sample plug-in. Continue with the next section where you modify this sample to include the custom visualization of the faceted data.

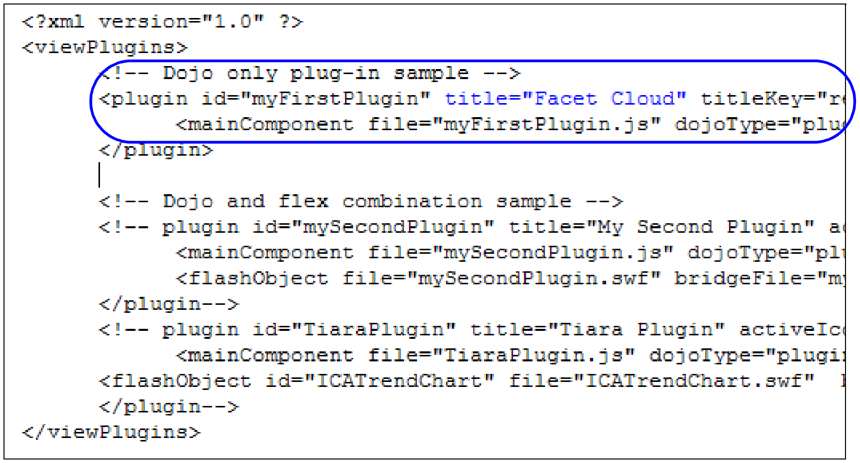
14.4 Customizing the sample text miner plug-in

This section shows how to add your own visualization to the myFirstPlugin sample. More specifically you add a *facet cloud*, which is a *word cloud* that shows the relative importance of the facet values based on their frequency counts.

Word clouds show more frequently used words in a larger font and sometimes with a different color. When viewing a word cloud, you can quickly see which words were used more frequently and, therefore, which words are potentially more important. Visually word clouds are thought to be quicker and more intuitive to comprehend rather than trying to decipher entries in a conventional list table.

14.4.1 Changing the view tab title

A simple modification is to change the title of the view tab. To change the title of the view tab, edit the `plugin.xml` file, which is the same file that was used to uncomment the myFirstPlugin sample. Change the title attribute from “My First Plugin” to “Facet Cloud” as highlighted in Figure 14-3.



```
<?xml version="1.0" ?>
<viewPlugins>
      <!-- Dojo only plug-in sample -->
      <plugin id="myFirstPlugin" title="Facet Cloud" titleKey="r
            <mainComponent file="myFirstPlugin.js" dojoType="plu
      </plugin>

      <!-- Dojo and flex combination sample -->
      <!-- plugin id="mySecondPlugin" title="My Second Plugin" a
            <mainComponent file="mySecondPlugin.js" dojoType="pl
            <flashObject file="mySecondPlugin.swf" bridgeFile="m
      </plugin-->
      <!-- plugin id="TiaraPlugin" title="Tiara Plugin" activeIc
            <mainComponent file="TiaraPlugin.js" dojoType="plugi
      <flashObject id="ICATrendChart" file="ICATrendChart.swf"  I
      </plugin-->
</viewPlugins>
```

Figure 14-3 Changing the title to a plug-in in the plugin.xml file

14.4.2 Customizing the plug-in template HTML file

Each plug-in has a JavaScript widget (`.js`) file and corresponding template HTML (`.html`) file. The template HTML file provides the overall layout of the view. The JavaScript widget references key elements of the template and populates those elements with data.

Edit the myFirstPlugin.html file in the ES_NODE_ROOT/master_config/searchapp/analytics/plugin/myFirstPlugin/templates directory. Example 14-1 shows the edited file.

Example 14-1 Updated HTML template with the facet cloud table added

```
<div style="width: 100%; padding: 5px; overflow: auto;" >
   <div style="font-size: large;">Cloud for selected facet:
        <span dojoAttachPoint="selectedFacetSpan"
       style="color: blue;"></span>
   </div>
   <table style="width: 90%; margin-top: 2em; border: solid black
1px;">
      <tbody dojoAttachPoint="cloudBody"></tbody>
   </table>
   <table style="width: 90%; margin-top: 2em; border: solid black
1px;">
      <thead style="background-color: #D4D0C8; font-weight: bold;">
         <tr>
            <th>Facet name</th>
            <th>Frequency</th>
            <th>Correlation</th>
         </tr>

      </thead>
      <tbody dojoAttachPoint="tableBody">
      </tbody>
   </table>
</div>
```

In addition to minor changes and additions in style specifications, notice that a new table was added before the conventional facet list table. The body of the facet cloud table is assigned a dojoAttachPoint of “cloudBody.” This name indicates how and where the facet cloud table will be referenced and updated with data by the myFirstPlugin.js JavaScript widget.

14.4.3 Customizing the javascript widget

The myFirstPlugin JavaScript widget does all of the work. It updates the HTML template file with data in response to certain events.

Update the `myFirstPlugin.js` file in the `ES_NODE_ROOT/master_config/searchapp/analytics/plugin/myFirstPlugin` directory:

1. Add the function shown in Example 14-2 to the JavaScript file.
2. In the _onLoad function, insert a line to call the function as shown in Example 14-2.

Example 14-2 The renderFacetCloud JavaScript function

```
_renderFacetCloud: function(facetValues) {

   dojo.empty(this.cloudBody);
   var maxFacets = 75;// Maxiumum number of facets values to display in the cloud
   var htmlCloud = "";// Contains the html for the cloud (to be incrementally built)

   if (maxFacets > facetValues.length) maxFacets = facetValues.length;
   var minValue = facetValues[maxFacets-1];
   var offset = minValue.weight-1;
   var maxValue = facetValues[0];

   // Divide the facet frequencies into groups of 7 (seven increasing font sizes)
   var factor = Math.round((maxValue.weight - minValue.weight + 1) / 7);
   if (factor == 0) factor = 1;

   var rands = [];// Array to hold the number of random numbers generated
   var numRands = 0;// Current number of random numbers generated


   for (var i=0;i<maxFacets;i++) {

       // Randomly select a facet that has not been selected yet
       var rand=Math.floor(Math.random()*maxFacets);
       while (true) {
      var j = 0;
      for (;j<numRands;j++) {
          if (rand == rands[j]) {
         rand=Math.floor(Math.random()*maxFacets);
         break;
          }
      }
      if (j >= numRands) {
         rands[j] = rand;
         numRands = numRands + 1;
         break;
      }
       }
```

```
        // Add the randomly select facet value to our cloud and set
        // its span class font size dependent on its frequency (weight)
        var facetValue = facetValues[rand];
        var fontsize = Math.round((facetValue.weight - offset) / factor);
        if (fontsize <=0) fontsize = 1;
        if (fontsize > 7) fontsize = 7;
        htmlCloud = htmlCloud + "<SPAN class=\"topicCloudSize"+fontsize+"\">";
        htmlCloud = htmlCloud + facetValue.label;
        htmlCloud = htmlCloud + " </SPAN>&nbsp;";
    }

    // Add our html cloud as a cell in the cloud table (body)
         var tr = dojo.create("tr", null, this.cloudBody);
         dojo.create("td", {innerHTML: htmlCloud}, tr);
},
```

3. In the _onLoad function, after the facet values have been retrieved, insert a call to the renderFacetCloud function (Example 14-3).

Example 14-3 Inserting the call to the _renderFacetCloud function

```
_onLoad: function(data) {
    if(data && data["es:apiResponse"] && data["es:apiResponse"]["ibmsc:facet"]) {
           var facetValues =
data["es:apiResponse"]["ibmsc:facet"]["ibmsc:facetValue"];
      this._renderFacetCloud(facetValues);
```

4. Increase the number of facet values that will be returned from the default of 10 to 100. Do this step near the top of the file when defining the variable for the facet string. Insert the `count` parameter set to 100 as shown in Example 14-4.

Example 14-4 Increasing the count of facet values returned

```
var _facet = {
        "id": facetId,
      "count": "100", // We want 100 maximum facet values returned instead of the
default of 10
        "namespace": (facetType == "subcategory" ? "keyword" : facetType)// show
keywords if subcategory
        };
```

5. Save the `myFirstPlugin.js` file.

14.4.4 Updating the style sheet for the plug-in

In the _renderFacetCloud JavaScript function, you might have noticed the use of SPAN tags around the facet values because they were added to the cloud. Each SPAN tag refers to one of seven style classes that control the size and color of the font for the facet value. The last task is to add those styles to the `common.css` file for the text miner application.

1. Edit the `common.css` file in the `ES_INSTALL_ROOT/jetty/searchapp/analytics/` directory. At the end of the file, add the entries shown in Example 14-5.

Example 14-5 Adding facetCloud styles to the common.css file

```
/**********************/
/* Topic Cloud Styles */
/**********************/
.topicCloudSize1 {
   font-size: 9pt;
   color: #b4b4b4;;
   font-family: Arial
}
.topicCloudSize2 {
   font-size: 12pt;
   color: #6E96C8;
   font-family: Arial
}
.topicCloudSize3 {
   font-size: 14pt;
   color: #85BEE7;
   font-family: Arial
}
.topicCloudSize4 {
   font-size: 16pt;
   /* color: #5b00b7; */
   color: #3EE9DB;
   font-family: Arial
}
.topicCloudSize5 {
   font-size: 18pt;
   /* color: #400040; */
   color: #41B041;
   font-family: Arial
}
.topicCloudSize6 {
   font-size: 20pt;
   color: blue;
   font-family: Arial
}
.topicCloudSize7 {
   font-size: 22pt;
   color: #dd00dd;
   font-family: Arial
}
```

2. Save the file.

14.5 Testing the customized plug-in

To test the changes made to the sample plug-in, follow these steps:

1. Restart the text miner application.
   – If you use the provided Jetty web server, enter the following commands, where *node_ID* identifies the search server:

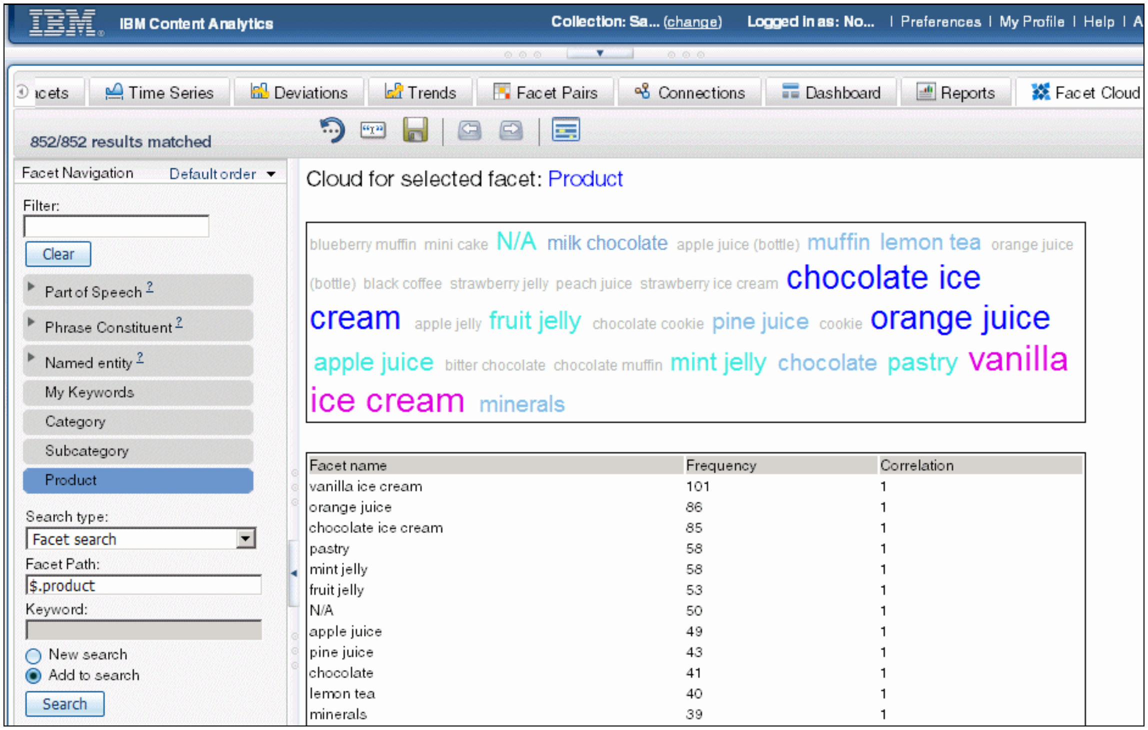
     ```
     esadmin session searchapp.node_ID restart
     ```

   – If you use WebSphere Application Server, enter the following command:

     ```
     esadmin config sync
     ```

2. Stop and restart the text miner application.

You then see the customized plug-in with a facet cloud being displayed as shown in the example in Figure 14-4.



852/852 results matched

Filter:

Search type: Facet search

Facet Path: $.product

Keyword:

○ New search
● Add to search

Cloud for selected facet: Product

blueberry muffin mini cake N/A milk chocolate apple juice (bottle) muffin lemon tea orange juice (bottle) black coffee strawberry jelly peach juice strawberry ice cream chocolate ice cream apple jelly fruit jelly chocolate cookie pine juice cookie orange juice apple juice bitter chocolate chocolate muffin mint jelly chocolate pastry vanilla ice cream minerals

Facet name	Frequency	Correlation
vanilla ice cream	101	1
orange juice	86	1
chocolate ice cream	85	1
pastry	58	1
mint jelly	58	1
fruit jelly	53	1
N/A	50	1
apple juice	49	1
pine juice	43	1
chocolate	41	1
lemon tea	40	1
minerals	39	1

Figure 14-4 Custom plug-in shown a facet cloud

15

Performance tips

IBM Content Analytics is an enterprise-level product that includes many components and complex architecture. This chapter outlines basic tuning tips for the main components of Content Analytics to ensure optimal system performance. It provides details about the different techniques for monitor and identifies potential bottlenecks in the system.

This chapter also includes hints and basic configuration for tuning the system. For capacity planning, contact your IBM representative.

This chapter includes the following sections:

- General performance guidelines
- Tuning the crawler component
- Tuning the document processor
- Tuning the indexer
- Enhancing the search performance
- Scalability
- Monitoring the system

15.1 General performance guidelines

Performance of the system is a result of several factors, including improper tuning of the product, the type of hardware, misconfiguration of the system, and network connectivity. Besides the hardware and software configuration, the type of data and load of the system can influence the performance of the system.

This section provides a starting point for tuning Content Analytics. It provides information about the factors to consider when setting parameters and defines the variables that are used later in the configuration. Later this chapter explains how to tune the different components of the system. It also provides ways to monitor the system to do quick problem analysis.

The basic tuning strategy of Content Analytics is similar to the general tuning strategy of any server. You monitor the processor utilization, I/O frequency, and throughput. Then you balance these parameters by removing any bottlenecks by following guidelines and tips provided in this chapter.

15.1.1 Factors that influence the performance of the system

Content Analytics has many components and different factors that can affect the performance of these components. In most cases, crawler does not need much tuning. However, indexer and document processors are resource consuming services. They require high performing machines and tuning to achieve good performance. Similarly, collections with a large number of data and facets can degrade the search performance. Before you tune the Content Analytics system, you must understand how the factors in the following sections can affect the components of the system.

Number of collections

Collections have a set of crawlers, an indexer, document processors, and search servers. They consume memory and processor resources when they are running. You can selectively stop collections to conserve resources. The more collections you have in the system, the more resources are required. Thus, to make resource management simple, use a single collection at a time.

Number of documents per collection

The amount of required resources relies on the total byte size of the ingested documents. The number of documents per collection is used to estimate the total size of the ingested document set.

Average document size

The average document size is used to estimate the total size of the document set. This number is used to estimate resources that are required by the document processor component. Also, rich text formats, such as PDF and Microsoft Word documents, often contain a smaller number of bytes of document text than the byte sizes of the files on disk. However, such documents require extraction of plain text from the binary data. Thus document processors often require more processor power and memory for such rich-text documents.

Largest document size

A document processor generally consumes 100 times the memory space of the size of document that is being processed. To prevent a shortage of memory, determine the maximum heap size of the document processors based on the size of the largest document in the document set to be ingested. The extracted document text that exceeds 128K characters is truncated in text analytics collections.

Available time window for rebuilding the index

When more data is added to the collection or when configuration changes are made, you must rebuild the index. Rebuilding the index is a time consuming task. It is usually scheduled for downtime during the day, which puts a constraint on the amount of time allocated to rebuild the index. Thus, to plan a consistent rescheduling, an acceptable hardware configuration is required.

Number of users

Content Analytics can serve up to 20 users with a high performance search server. Add additional search servers if you expect more than 20 users to access the system concurrently.

Additional configurations

Aside from the previously mentioned factors, the following additional features can also affect the performance of the system:

- Enabling secure search
- Using static ranking
- Using thumbnail generation
- Using global processing
- Exporting analyzed documents
- Changing facet generation rules and field definitions
- Adding user dictionaries, user patterns, named entity extraction rules, IBM Classification Module annotator, or custom annotators

15.1.2 Variables

Many sections of this chapter explain how to set the parameters for performance tuning in Content Analytics. This section describes the common variables that are used in those steps.

ES_INSTALL_ROOT

ES_INSTALL_ROOT is an environment variable that indicates the location of the Content Analytics installation. The default value for this variable is `C:\Program Files\IBM\es` on Windows operating systems and `/opt/IBM/es` on UNIX operating systems.

ES_NODE_ROOT

ES_NODE_ROOT is an environment variable that indicates the location of user data. This variable is set by the installation wizard. The default value of this variable is the `esdata` directory under the home directory of the user that you specify during the installation. On the master server, this directory contains the master configuration files. It is an important directory in regard to the tuning parameters.

Collection ID

The collection_ID variable is the ID of the collection for which you want to set a parameter value. This name is not the collection name that you specified in the first field when you created the collection.

You can specify the collection ID in the advanced options when you create the collection. For example, we specified `col_sample` as the collection ID in Chapter 4, “Installing and configuring IBM Content Analytics” on page 71. Otherwise, the system automatically assigns the ID a random number, such as `col_12345`.

To determine the collection ID after a collection is created, click **View collection settings** on the **General** page of the Collections view in the administration console. See “Hints and tips for using the esadmin utility” on page 599 to determine the collection ID by using the command-based utility.

Node ID

The node_ID variable is the ID of the server. The master server is assigned the ID node1, and additional servers are assigned IDs, such as `node2` and `node3`, in the order in which they are added to the system.

Session ID

The session_ID variable is based on the collection ID, node ID, and the component name as in the following example:

```
Indexer session ID: collection_ID.indexservice
Document processor session ID: collection_ID.docproc.node_ID
Search server session ID: collection_ID.searcher.node_ID
```

You can determine the session IDs by entering the **esadmin check** command at the command prompt shown in Example 15-1.

Example 15-1 Sample processor IDs

```
bash-3.2$ esadmin check
Session ID             Node ID     PID   State
------------------------------------------------------
col_sample.docproc.node1        node1       -     -
col_sample.exporter             node1       -     -
col_sample.indexservice         node1       -     -
col_sample.searcher.node1       node1       -     -
col_sample.stellent             node1       -     -
...
```

A session indicates an instance of a component. Thus, if you have an additional search node in the environment, you see the `col_sample.searcher.node2` session in the list. See “Hints and tips for using the esadmin utility” on page 599 for further details about using the **esadmin** utility to find session IDs.

Session configuration ID

The config_ID variable is the ID of the session configuration that is automatically assigned in the session configuration file in the `$ES_NODE_ROOT/master_config/collection_ID_config.ini` directory. Example 15-2 shows a sample of a session configuration.

Example 15-2 Sample configuration

```
session1.clone_sid=col_sample.indexservice
session1.collectionid=col_sample
session1.configDir=col_sample.indexservice
session1.dataDir=
session1.description=Index Document Processing Session
session1.displayname=Index Document Processor - Collection a (node1)
session1.domain=.
session1.flags=0
session1.hard_max_heap=10,8192
session1.id=col_sample.docproc.node1
```

```
session1.max_heap=8192
session1.nodeid=node1
session1.sectiontype=session
session1.subtype=docproc
session1.type=indexservice
...
```

You must know how to find the variables because they are used when making configuration changes later in this chapter.

15.2 Tuning the crawler component

Content Analytics can collect and extract content from various data source types. The performance characteristics of crawlers vary depending on the data source. The data sources often become a bottleneck if they are not appropriately tuned for crawling. Configuring and tuning the data source system is outside the scope of this book.

In a case where a crawler is a bottleneck, you can improve the performance by increasing the crawler instance, increasing crawler threads, or setting the appropriate memory. For example, a common characteristic among crawlers is the use of a database. When crawling data sources incrementally, the crawlers store information, such as visited locations, crawled time, and last modified date, in the database tables. The size of the database tables usually increases as you increase the number of retrieved documents. When you retrieve a large number of documents, the scanning of the database record becomes the bottleneck.

To prevent this situation, create multiple crawlers to separate the crawl space into multiple crawl spaces containing about 10 million documents. For example, you have a file system as a data source that has eight directories, containing 5 million documents each (40 million total). In this case, it is better to have four file system crawlers that crawl two directories (10 million each), instead of a single file system crawler to retrieve whole 40 million documents.

This section focuses on how to configure a crawler for optimal performance.

15.2.1 Increasing active crawler threads

The default value for the maximum number of active crawler threads is 10. In an environment where data source is tuned for performance, you can increase the maximum number of crawler threads to improve the performance of the crawler.

To adjust the maximum number of active crawler threads, change the number of indexer threads:

1. Log in to the administration console.
2. Click **Collections** in the toolbar to open the Collections view.
3. In the list of collections, locate the collection that you want to edit, and click the **Edit** icon (Figure 15-1).

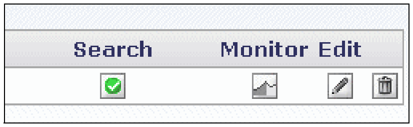


Figure 15-1 Edit icon in the collections view

4. Click the **Crawler** tab, and click **Crawler Properties** (Figure 15-2).

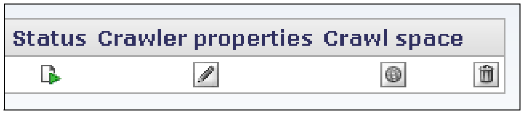


Figure 15-2 Editing the crawler properties

5. Click the **Advanced options** link (Figure 15-3).

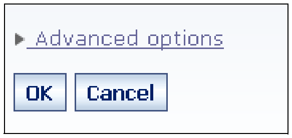
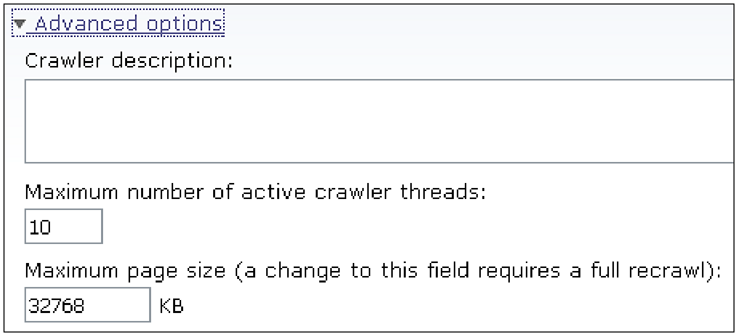


Figure 15-3 Advanced options for crawler

6. In the Maximum number of active crawler threads field, modify the value and click **OK** (Figure 15-4).



Crawler description:

Maximum number of active crawler threads: 10

Maximum page size (a change to this field requires a full recrawl): 32768 KB

Figure 15-4 Editing the number of crawler threads

7. Restart the crawler services

Real-time monitor: Content Analytics provides a real-time monitor for web crawler with such statistics as retrieval rate and active sites. In this case, web crawler is performing optimally if the crawler rate is similar to the number of active threads and number of active sites.

15.2.2 Setting the maximum heap size

You can create multiple crawlers in a collection. These crawlers use computing resources only when they are running. To determine the required memory size of the crawler, add 100 MB for every concurrent crawler instance, except for web crawlers. Memory consumption of a web crawler instance can be up to 500 MB. Because all crawlers are created on a single server (a master server or a crawler server), you must use a high performance computer if you plan to run many crawler instances at the same time. By default, the system sets their memory sizes to 512 MB.

To change the maximum heap size of a crawler component, follow these steps:

1. Stop the system:

```
esadmin stop
```

2. Stop the common communication layer (CCL) server:

```
stopccl
```

3. In the `$ES_NODE_ROOT/master_config/collection_ID_config.ini` file, find the session configuration for which you want to change the heap size. The session configuration includes the following lines:

   ```
   config_ID.id=session_ID
   config_ID.id=crawler_name
   config_ID.max_heap=max_heap_size_in_MB
   ```

   For example, the following lines are from a session configuration of a crawler session for a collection with the id `col_sample`:

   ```
   session1.id= col_sample.WIN_7776
   session1.displayname=IVPFile System Windows file system
   session1.max_heap=512
   ```

4. Change the value of the max_heap property, and save the file.
5. Restart the CCL server:

   ```
   startccl
   ```

6. Restart the system:

   ```
   esadmin start
   ```

15.3 Tuning the document processor

The document processor parses the document to extract text and metadata. It also tokenizes the extracted text data so that the text can be indexed. Unstructured Information Management Architecture (UIMA) pipelines are also processed in the document processors, and a large amount of annotations is generated.

By default, a document processor is created on a master server so that you can build an index before you add additional document processor servers. After adding one or more additional document processor servers, you can stop the document processor on the master server to allocate more computing resources for other components. You can also create multiple instances (threads) of a document processor to further improve the document processing power.

Also this component can be scaled-out to multiple additional servers. The number of document processors is logically not limited. However, the effective number of document processors is often limited by the throughput of the indexer, which depends on the computing power of the master server.

The focus of this section is on how you can configure a document processor for optimal performance.

15.3.1 Setting the number of document processor threads

The default number of document processors for each server is 4. If you use powerful servers for the additional document processor servers, you might want to increase the number of document processor instances for each server to use the many cores. Set the number of document processors equal to 2 times the number of processor cores of document processor servers.

To change number of document processor threads, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID.indexservice/collection.properties` file, modify the NumberOfDocumentProcessors property. If you need to increase the number of stellent parsers for rich text documents, change the value of the NumberOfStellentParsers property to the same value as the NumberOfDocumentProcessors property.
4. In the `$ES_NODE_ROOT/master_config/collection_ID.indexservice/collection.xml` file, find the <tokenizer> element under the <config> and <collection> elements.
5. Add the <numberOfTokenizers> element in the <tokenizer> element:

   ```
   <tokenizer>
   <ngramMode>hybrid</ngramMode>
   <numberOfTokenizers>8</numberOfTokenizers>
   </tokenizer>
   ```

6. Modify the number of tokenizers. The number of document processors and the number of tokenizers must be the same.
7. Save the files.
8. Restart the CCL server:

   ```
   startccl
   ```

9. Restart the system:

   ```
   esadmin start
   ```

15.3.2 Increasing the maximum heap size

The required memory size for each document processor instance can be calculated from the size of the largest amount of textual data in the ingested documents. For the maximum heap size of each document processor instance, allocate 100 times more memory than the size of the largest document. In addition, add 100 MB for the dictionary data, and add another 200 MB if the named entity annotator is enabled to the maximum heap size. By default, the system sets the maximum heap size to 1024 MB.

To change maximum heap size of a document processor instance, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID_config.ini` file, find the session configuration for which you want to change the heap size. The session configuration includes the following lines:

   ```
   config_ID.id=session_ID
   config_ID.max_heap=max_heap_size_in_MB
   ```

   For example, the following lines are from a session configuration of an indexer session for a collection with the id `col_sample`:

   ```
   session3.id= col_sample.WIN_7776
   session3.max_heap=512
   ```

4. Change the value of the max_heap property, and save the file.
5. Restart the CCL server:

   ```
   startccl
   ```

6. Restart the system:

   ```
   esadmin start
   ```

15.4 Tuning the indexer

The indexer component creates the index based on the ingested documents. In most text analytics collections, this component can become a bottleneck because it is a single component that needs to process large amounts of data that are generated in a high throughput by numerous document processors. If you plan to use many document processors or want to improve the end-to-end index building process, use your most powerful server as the master server.

The indexer is also an I/O intensive task because it writes a large amount of data to the disk. Therefore, the number of disk I/Os often becomes the bottleneck of the indexing throughput. The indexer temporarily stores the data in the buffer residing in the memory. The data from the buffer is written to the disk either when the buffer is full or when the commit interval is met. Thus, a larger buffer size or longer commit interval reduces the number of disk I/Os and results in an increased total throughput.

Lastly, the index often requires rebuilding when changes are made to the configuration. You must tune the indexer appropriately to optimize rebuilding the index. This focus of this section is on how you can carefully tune an indexer for optimal performance.

15.4.1 Setting the number of indexer threads

To efficiently use many document processors, the number of indexer threads is important. The indexer threads delegate the ingested documents to the document processor. The threads wait (sleep) during the document processing until they receive the processed documents that are ready to be stored in the index. Thus you must also take into account the number of waiting (sleeping) indexer threads.

You can assume that the number of waiting indexer threads is the same as the number of document processors. The default number of indexer threads is 1. In most cases, you need to increase the number of threads. Typically, the recommended number of indexer threads is equal to 2 to 4 times the number of processor cores in the master server machine plus the total number of document processors:

```
Number_of_threads = (2 X number_of_CPU_cores_of_master_server) +
                     number_of_document_processor_threads
```

To change the number of indexer threads, follow these steps:

1. Log in to the administration console.
2. Click **Collections** in the toolbar.

3. In the Collections view, in the list of collections, find the collection that you want to edit, and click **Edit** (Figure 15-5).

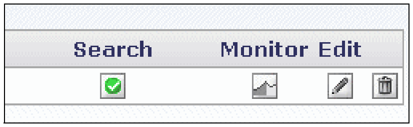


Figure 15-5 Clicking the Edit icon in the Collections view

4. Select the **Parse and Index** tab, and click **Configure parsing options** (Figure 15-6).

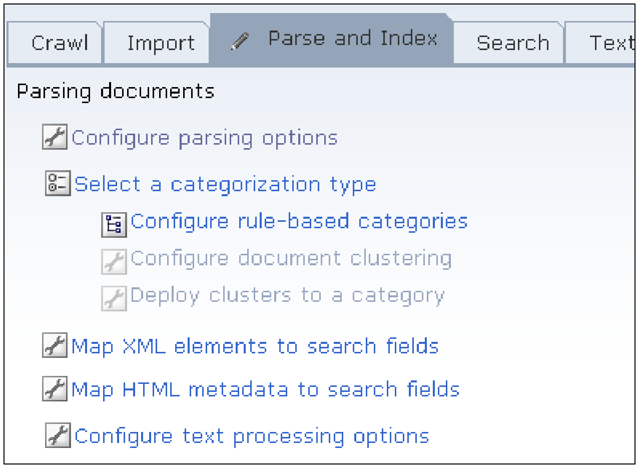
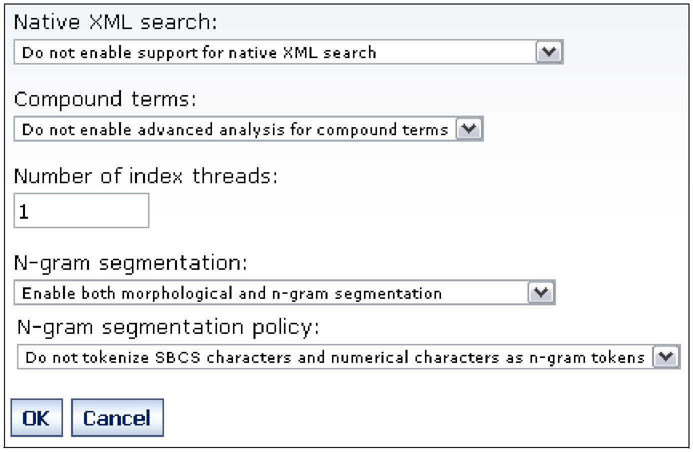

Crawl
Import
Parse and Index
Search
Text
Parsing documents
Configure parsing options
Select a categorization type
Configure rule-based categories
Configure document clustering
Deploy clusters to a category
Map XML elements to search fields
Map HTML metadata to search fields
Configure text processing options

Figure 15-6 Selecting Configure parsing options

5. In the Number of index threads field, modify the value and click **OK** (Figure 15-7).



Native XML search:
Do not enable support for native XML search

Compound terms:
Do not enable advanced analysis for compound terms

Number of index threads:
1

N-gram segmentation:
Enable both morphological and n-gram segmentation

N-gram segmentation policy:
Do not tokenize SBCS characters and numerical characters as n-gram tokens

Figure 15-7 Setting the number of index threads

6. Restart the parsing and indexing services.

15.4.2 Specifying the taxonomy cache type

The taxonomy cache stores generated facets that are used by the indexer component. Content Analytics provides three types of taxonomy caches:

LRU	Partial in memory cache where scalability is limited.
TrieL2O	Complete in memory cache that is 2-5 times faster than LRU.
DA	Enhancement of in-memory cache, provides more tuning parameters than TrieL2O.

By default, the LRU cache is used. If more memory is available, set the cache type as either TrieL2O or DA to fully store the facets in the memory and improve the overall indexing throughput. However, if you want to process a large document set in a longer duration with a smaller memory footprint, you can configure the system to use the LRU cache that partially loads the taxonomy index in memory.

To change the default taxonomy cache type, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID.indexservice/collection.xml` file, find the <index> element with a <type> element value of Facet. This element has the following XPath:

   ```
   /config/collection/indexes/index[type=Facet]
   ```

4. Add the CacheType property to the <index> element, and set its value to DA, TrieL2O, or LRU:

   ```
   <index>
   <type>Facet</type>
   <path>facets</path>
   <enabled>true</enabled>
   <indexMode mode="normal"/>
   <property name="CacheType" value="LRU"/>
   ...
   </index>
   ```

5. Save the file.
6. Restart the CCL server:

   ```
   startccl
   ```

7. Restart the system:

   ```
   esadmin start
   ```

15.4.3 Increasing the maximum heap size

When you estimate the required memory size of the indexer, it is important to know that the indexer has three types of caches that are used to process facets. The required memory highly depends on the type of cache set for the collection and the total size of the data.

In the default setting (in which the maximum heap size is 1024 MB or less), the indexer uses the specified size of memory to store the facets in memory (the LRU taxonomy cache) and requires approximately 20 MB. This memory consumption is included in the default maximum heap setting. When all facets are stored in memory (the TrieL2O and DA taxonomy cache), 250 MB is required for every 1 GB of document text.

In addition to the memory consumption of facets, estimate 50 MB for each indexer thread. If you plan to configure more than 10 threads for the indexer, add 500 MB as a safety margin for the buffers and other tasks of the indexer.

You can calculate the estimated maximum heap size by using following formula:

- For TrieL2O and DA:

  ```
  Required_Memory = (500 MB or 50 MB per thread) + 250 MB per 1 GB
  document text
  ```

- For LRU:

  ```
  Required_Memory = (500 MB or 50 MB per thread) + 50 MB per 1 GB
  document text
  ```

To change the maximum heap size of an indexer, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID_config.ini` file, find the session configuration for which you want to change the heap size. The session configuration includes the following lines:

   ```
   config_ID.id=session_ID
   config_ID.max_heap=max_heap_size_in_MB
   ```

   For example, the following lines are from a session configuration of an indexer session for a collection with the id `col_sample`:

   ```
   session3.id=col_sample.indexservice
   session3.max_heap=1024
   ```

4. Change the value of the max_heap property, and save the file.
5. Restart the CCL server:

   ```
   startccl
   ```

6. Restart the system:

   ```
   esadmin start
   ```

15.4.4 Increasing the buffer size

The buffer_size property determines the size of the temporary storage in the memory used by the Lucene index writer to store processed data before it is written to the index. The value of the buffer size is determined from the target indexing throughput. In most cases, assign 50 MB for every 1 million documents.

To change the buffer size, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID.indexservice/collection.xml` file, find the <index> element with a <type> element value of Text. This element has the following XPath:

   ```
   /config/collection/indexes/index[type=Text]
   ```

4. Find the parameter named BufferSize under the <index> element, and rewrite its value with your new value:

   ```
   <indexes>
         <index>
           <type>Text</type>
           <path>text</path>
           <enabled>true</enabled>
           <indexMode mode="normal"/>
           <property name="SweetSpotLength" value="50"/>
           <property name="BufferSize" value="2048"/>
   ```

5. Save the file.

6. Restart the CCL server:

   ```
   startccl
   ```

7. Restart the system:

   ```
   esadmin start
   ```

15.4.5 Increasing the index commit interval

A commit operation happens when data is written to the disk. A *commit interval* is a condition in which the commit operation occurs. To increase the interval of the index commit operation, set the values for the numberOfDocumentsTilFlush and DL_RDS_File_Limit parameters.

The first parameter, numberOfDocumentsTilFlush, specifies the number of ingested documents that triggers the index to perform the index commit operation. By setting a small value for this parameter, the ingested documents can be searchable after a short time, but the throughput is decreased. If maximum throughput is required, this parameter can be set as -1, which means that index commit is not triggered based on the number of ingested documents.

The commit interval in bytes must be smaller than one-fourth of the buffer size of the index. The recommended value for the numberOfDocumentsTilFlush parameter is calculated by using following formula:

```
numberOfDocumentsTilFlush = (indexBufferSize * numberOfPartitions / 4)
/ average_document_size
```

For example, consider the following example:

- indexBufferSize = 2048 MB
- numberOfPartitions = 1
- average_document_size = 4096 bytes

Based on these values, the numberOfDocumentsTilFlush is calculated as follows:

```
(2048 x 1024 x 1024 x ¼) / 4096 = 131,072 documents
```

To set the numberOfDocumentsTilFlush parameter, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID.indexservice/collection.xml` file, find the <indexer> element under the /config/collection element.

4. Add the <numberOfDocumentsTilFlush> element to the <indexer> element:

   ```
   <indexer>
     <numberOfDocumentsTilFlush>50000</numberOfDocumentsTilFlush>
   </indexer>
   ```

5. Save the file.

6. Restart the CCL server:

   ```
   startccl
   ```

7. Restart the system:

   ```
   esadmin start
   ```

The second parameter, DL_RDS_File_Limit, controls the maximum size of the RDS files. RDS files are used internally in the queue between the crawlers and the indexer. RDS files temporarily store the crawled documents on the disk to safely keep the crawled documents until they are written in the index. The Crawled documents are written in several chunked RDS files. Each RDS file is removed when all documents included in the file are successfully written in the index. Therefore, it is necessary to enlarge the size of the RDS files if you maximize the indexing throughput.

To set the maximum size of the RDS files, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. Open the `$ES_NODE_ROOT/master_config/datalistener/dlConfig.prp` file.
4. Add the DL_RDS_File_Limit property with the new size in KB:

   ```
   DL_RDS_File_Limit=1024
   ```

5. Save the file.
6. Restart the CCL server:

   ```
   startccl
   ```

7. Restart the system:

   ```
   esadmin start
   ```

15.5 Enhancing the search performance

The search server works with search applications to process queries and to search the index to provide the search and analytics capability to users. The search server can be scaled out to multiple additional servers.

Content Analytics provides two index sharing alternatives. The first way is to *copy and share nothing* between the search servers. In this case, each additional search server requires sufficient disk space for a replica of the index. The replica includes the main text index and the thumbnails, which require the majority of the disk space.

The second way (the recommended way) is to *share all*. This way entails using the General Parallel File System (GPFS™) in a storage area network (SAN), with the enterprise class storage (higher than the IBM System Storage DS4000®

series). You can set up index sharing by specifying the GPFS shared directory as the data directory at installation time. You need a load balancer to distribute the requests of the users to the multiple search servers.

Several factors, such as the number of query requests, result for the query, number of facets, and how they are loaded into the memory, influence how the search server performs. You can enhance the performance of the search server by setting a large heap size, a search cache, and document pooling. This section focuses on how to configure a search server for optimal performance.

15.5.1 Increasing the search result cache entries

When a query is executed and results are returned, Content Analytics first evaluates the result cache entries to see if results can be returned to the user from the cache. Increasing the size of the cache enhances the search performances by decreasing the response time of the query. You can set the maximum number of entries that are saved in the cache as explained in the following steps:

1. Log in to the administration console, and click **Collections** in the toolbar.
2. In the Collections view, in the list of collections, locate the collection that you want to edit, and click **Edit** (Figure 15-8).

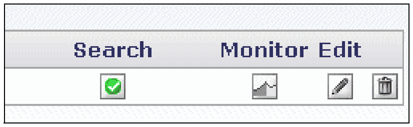


Figure 15-8 Clicking the Edit icon in the Collections view

3. Click the **Search** tab (Figure 15-9) and click **Configure search server options**.

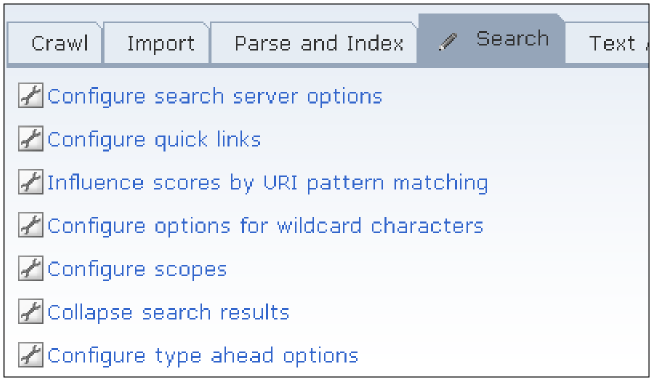
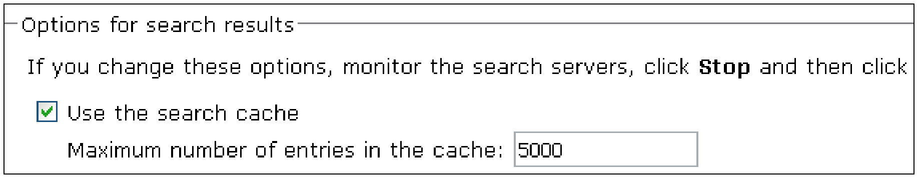

Crawl
Import
Parse and Index
Search
Text
Configure search server options
Configure quick links
Influence scores by URI pattern matching
Configure options for wildcard characters
Configure scopes
Collapse search results
Configure type ahead options

Figure 15-9 Selecting to configure the search server options

4. Under Options for search results (Figure 15-10), ensure that the **Use the search cache** check box is selected, and modify the value in the “Maximum number of entries in the cache” field. The server cache can store up to 5000 entries. Click **OK** to save changes.



Options for search results

If you change these options, monitor the search servers, click **Stop** and then click

☑ Use the search cache

Maximum number of entries in the cache: 5000

Figure 15-10 Enabling the search cache

5. Restart the search server.

15.5.2 Enabling the optional facet index

The default mechanism for querying facets might perform slowly if you analyze a collection that contains more than 5 million documents. In such cases, enable the optional facet index, as explained in the following steps, to optimize the facet counting algorithm. A long building time is usually required for the optional facet index, especially if the index includes a large number of facets. The amount of time that is required to build the optional facet index is almost the same as the amount of time that is required to build the main index.

To enable the optional facet index, follow these steps:

1. Log in to the administration console, and click **Collections** in the toolbar.
2. In the Collections view (Figure 15-11), in the list of collections, locate the collection that you want to edit, and click **Edit**.

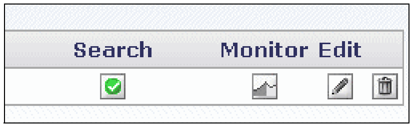


Figure 15-11 The Edit option in the Collections view

3. Select the **General** tab, and click **Configure general options** (Figure 15-12).

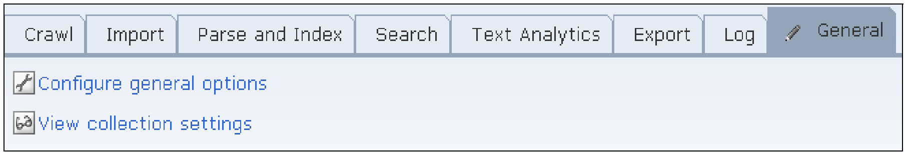
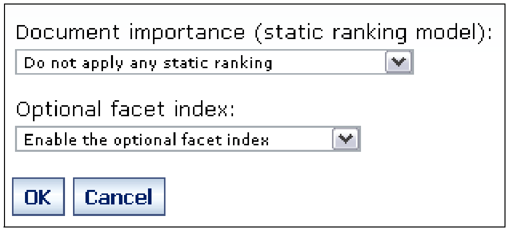


Figure 15-12 Selecting Configure general options

4. For Optional facet index (Figure 15-13), select **Enable the optional facet index**. Click **OK** to save the changes.

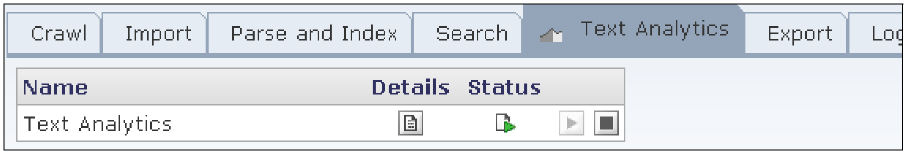


Document importance (static ranking model):
Do not apply any static ranking

Optional facet index:
Enable the optional facet index

Figure 15-13 Enabling the optional facet index

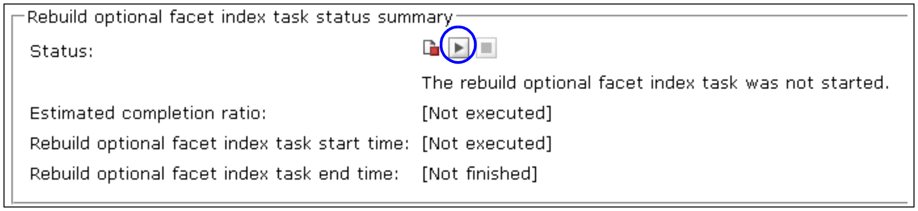
5. Click the **Monitor** icon.
6. Select the **Text Analytics** tab, and click the **Details** icon (Figure 15-14).



Name	Details	Status
Text Analytics		

Figure 15-14 Monitoring Text Analytics

7. Build the optional facet index (Figure 15-15).


Rebuild optional facet index task status summary

Status:	The rebuild optional facet index task was not started.
Estimated completion ratio:	[Not executed]
Rebuild optional facet index task start time:	[Not executed]
Rebuild optional facet index task end time:	[Not finished]

Figure 15-15 Building the optional facet index

15.5.3 Increasing the maximum heap size

If you are analyzing large collections, you typically have a large number of document text and facets that require configuring the maximum heap size of the search servers. For example, search servers store facet data for all documents in memory. Thus, the performance of facet querying depends on the number of documents and the number of facets. To store facet data in memory, add 150 MB to the maximum heap size of the search server component for every 1 GB of document text.

You can determine the size of the document text that is created in the following directory on the master server:

- In UNIX, the `$ES_NODE_ROOT/data/Collection_name/index/text` directory
- In Windows, the `%ES_NODE_ROOT%\data\Collection_name\index\text` directory

Additionally, if search result cache feature is enabled, adjust the heap size accordingly. On average, the default 3000 entries of search result cache consumes about 100 MB heap.

To change maximum heap size of a search instance, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID_config.ini` file, find the session configuration for which you want to change the heap size. The session configuration includes the following lines:

   ```
   config_ID.id=session_ID
   config_ID.max_heap=max_heap_size_in_MB
   ```

For example, the following lines are from a session configuration of a searcher session for a collection with the id `col_sample`:

```
session6.id=col_sample.searcher.node1
session6.max_heap=4096
```

4. Change the value of the max_heap property and save the file.
5. Restart the CCL server:

   ```
   startccl
   ```

6. Restart the system:

   ```
   esadmin start
   ```

15.5.4 Setting the number of threads for rebuilding optional facet index

When you build the optional facet index with many cores, you might want to tune the concurrency of this task. Basically the number of threads must be 2-4 times larger than the number of cores (depending on the capability of the running threads on each core).

Rebuilding the optional facet index includes a sorting phase and a writing phase. The writing phase is time consuming and mainly depends on the disk performance. You can only tune the sorting phase by using this parameter.

To set the number of threads for rebuilding optional facet index, follow these steps:

1. Stop the system:

   ```
   esadmin stop
   ```

2. Stop the CCL server:

   ```
   stopccl
   ```

3. In the `$ES_NODE_ROOT/master_config/collection_ID.indexservice/collection.xml` file find the <index> element with a <type> element value of Facet2. The XPath of this element is `/config/collection/indexes/index[type=Facet2]`.
4. Add the RebuildThreadCount property to the <index> element with the number of threads as its value:

   ```
   <index>
   <type>Facet2</type>
   <path>facets2</path>
   <enabled>true</enabled>
   ...
   ```

```
<property name="RebuildThreadCount" value="10"/>
</index>
```

5. Save the file.
6. Restart the CCL server:

   ```
   startccl
   ```

7. Restart the system:

   ```
   esadmin start
   ```

15.6 Scalability

If you have applied the guidelines mentioned in this chapter and are still experiencing undesired performance, consider scaling the system by adding more nodes. With Content Analytics, you can expand the configuration on multiple machines so that you can run search servers or document processor servers on additional nodes.

To install and configure additional nodes or distributed system, go to the IBM Content Analytics Information Center at the following address, and search on *installing additional nodes*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

15.7 Monitoring the system

Typically when you consider the performance, you might want to check the following information:

- The system resources, such as processor, memory, or disk I/O, are used as expected, and the product is running with acceptable performance. Alternatively, the system resources are not used well, and the product is not running with the acceptable performance (that is, slow performance).
- The product is up and running, but its speed is slow, or it has stopped processing (that is, hanging).

This section explains how to monitor the system by using either the Content Analytics commands-based utility (the **esadmin** command) or the administration console. (This section does not explain how to analyze the output of the commands or utilities, which is outside the scope of this book.) This section provides tips for using the operating system utility.

15.7.1 General guideline for monitoring the system

Generally speaking, when you consider performance tuning or troubleshooting, first you monitor the entire system. You must know how the system resources that are used. After you see that the system resources are not used as expected compared to the usual state, look further for any bottlenecks of the system resources such as processors, memory, and disk I/O.

In addition, when you consider performance tuning, you must know the usual system resource state to compare it to the abnormal system resource state. For example, if processor utilization is always 80%, which is a normal state to the system, consider adding more processors to the system to acquire the desired performance. If the usual system processor utilization is around 50% and is not used as much, consider what is blocking the processor from having a much higher utilization. Then look into the other resources, such as memory or disk I/O, and see whether a bottleneck is the cause. Also, if the usual system processor utilization is around 50%, but suddenly increases to 80%, you might want to investigate the reason why this change has occurred.

Further, when you monitor the system, you must perform the same operation periodically. That is, to confirm the process progress of the system, you must capture the system state by using commands several times with a certain interval that depends on the situation.

For example, consider the situation when you see that the crawler is extremely slow. For example, it does not finish crawling within an hour, when it usually finishes within 15 minutes. You notice that the current situation is extremely slow compared to the usual situation. In such case, you might want to confirm if the indicator, such as progress percentage, is continuously increasing or is not increasing every 5 minutes.

Thus, monitoring the operating system resources is important when you consider the performance tuning. It is also important for performance-related troubleshooting.

15.7.2 Using the esadmin command utility

Sometimes you might want to monitor the progress of processing Content Analytics. You can use the administration console to monitor the progress, but Content Analytics provides the **esadmin** command utility to monitor the progress.

This section explains how to use the **esadmin** command utility. However, note that the **esadmin** command for monitoring is not available for some components. You must still use the administration console in those cases.

> **Command output:** The **`esadmin`** command output is in the XML format on the same console where you issue the command. Therefore, you might need to redirect the output to the file if necessary. Also, the XML format might not be intuitive to understand sometimes. See the IBM Content Analytics Information Center at the following address, for further details, such as what the content of each element means:
>
> `http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

Monitoring the crawler status

When you consider monitoring the crawler status, be aware of the crawler that you use (the web crawler or the data source crawlers other than the web crawler), because the command option is slightly different. Therefore, you must use the proper option to monitor the crawler status depending on the crawler that you use.

When you monitor the crawler status, use the `getCrawlerStatus` option with a crawler session ID that you want to monitor as shown in the following example:

```
esadmin crawler_session_id getCrawlStatus
```

This command returns the result with the XML document format (shown in Example 15-3) when you specify the Windows file system crawler to monitor.

Example 15-3 Result of the esadmin getCrawlerStatus command

```
>esadmin col_sample.WIN_77148 getCrawlerStatus
FFQC5303I IVPFile System  (sid: col_sample.WIN_77148) CCL session
exists. PID: 3484
FFQC5314I The following result occurred: <?xml version='1.0'
encoding='UTF-8'?>
<GeneralStatus>
        <Status>1</Status>
        <StatusMessage>Running</StatusMessage>
        <NumberOfServers>1</NumberOfServers>
        <NumberOfCompletedServers>0</NumberOfCompletedServers>
        <NumberOfTargets>1</NumberOfTargets>
        <NumberOfCompletedTargets>0</NumberOfCompletedTargets>
        <NumberOfCrawledRecords>599</NumberOfCrawledRecords>
        <RunningThreads>1</RunningThreads>
        <Progress>70</Progress>
</GeneralStatus>
```

As you can see, the content of the `StatusMessage` element indicates that the crawler is running. Also, the entire progress is 70% as indicated by the `Progress` element. You can monitor the progress whether it is stuck at a certain point or proceeding.

Monitoring the parser status

Monitoring the parser status is straight forward. You can use `getCollectionParserMonitorStatus` option with the `collection ID` that you want to monitor as shown in the following example:

```
esadmin monitor getCollectionParserMonitorStatus -cid collection_ID
```

This command returns the result with XML document format as shown in Example 15-4.

Example 15-4 Result of the esadmin getCollectionParserMonitorStatus command

```
>esadmin monitor getCollectionParserMonitorStatus -cid col_sample
FFQC5303I Monitor (node1) (sid: monitor) CCL session exists. PID: 3996
FFQC5314I The following result occurred: <Monitor Type="Parser">
<ParserStatus><Status>1</Status></ParserStatus></Monitor>
```

This command only indicates whether the parser is up and running. If the content of the `Status` element is 1, the parser is running.

If you need to see how many documents are in process, watch the administration console that is updated automatically and see whether the number of documents processed is increasing.

Monitoring the search server status

Monitoring the search server status is also straight forward. You can view the basic status of search server through the administration console by viewing the search session details. For detailed information, you can use the **`esadmin`** command with the `monitor getCollectionSearchMonitorStatus` option and collection ID that you want to monitor as shown in the following example:

```
esadmin monitor getCollectionSearchMonitorStatus -cid collection_id
```

This command returns the result with the XML document format as shown in Example 15-5. In this example, we check the search server status for collection ID *`col_sample`*. Usually this command is useful for confirming if the search server is up and running. As you might imagine, you cannot analyze the data in the collection if the search server of the collection is not up and running.

Example 15-5 The esadmin monitor getCollectionSearchMonitorStatus command result

```
>esadmin monitor getCollectionSearchMonitorStatus -cid col_sample
FFQC5303I Monitor (node1) (sid: monitor) CCL session exists. PID: 3996
FFQC5314I The following result occurred: <?xml version="1.0"?>
```

```
<Monitor Type="Search" Count="1">
<SearchStatus Name="Search Manager (node1)"
SearchID="searchmanager.node1" HostName="cca.imecm.local">
<Status>1</Status>
<CacheHits>98</CacheHits>
<CacheHitRate>0.9423076923076923</CacheHitRate>
<QueryRate>0</QueryRate>
<ResponseTime>0</ResponseTime>
</SearchStatus>
</Monitor>
```

For further information about each command and the meaning of each element in the command result output, go to the IBM Content Analytics Information Center at the following address, and search on *commands, return codes, and session IDs*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Hints and tips for using the esadmin utility

When you use the **`esadmin`** utility, consider the collection ID for the collection in question. You can use the **`esadmin report collections`** command to help you see which collection ID is associated with the collection name, as shown in Example 15-6. In this example, you see that two collections and the collection ID (referred as `col_id`) correspond to `Display Name`.

Example 15-6 The result of the esadmin report collections command

```
>esadmin report collections
FFQC5320I
Collections Report
-----------------------------------------------
col_id:            col_sample
Display Name:      Sample Text Analytics Collection
Description:       Collection created by SIAPI client
Index Type:        3
Flags:             32
Data Directory:    C:\IBM\es\esadmin\data\col_sample
Index Data dir:    C:\IBM\es\esadmin\data\col_sample\indexbuild
Index config file: col_sample_config.ini
 -------- End of Report. Total: 1--------
```

Also, when you want to confirm the crawler session ID, the **esadmin report sessions** command generates the list of session IDs as shown in Example 15-7. You can confirm the collection ID by using the **esadmin report collections** command first and seeing if the crawler is defined.

Example 15-7 Result of the esadmin report sessions command

```
>esadmin report sessions
Session ID                                 Node ID
----------------------------------------------------------------------
admin                                      node1
col_sample.WIN_77148                       node1
col_sample.WIN_77148.crawlerplugin         node1
col_sample.docproc.node1                   node1
col_sample.exporter                        node1
col_sample.indexservice                    node1
col_sample.searcher.node1                  node1
col_sample.stellent                        node1
configmanager                              node1
controller                                 node1
converter.node1                            node1
customcommunication                        node1
database                                   node1
datalistener                               node1
discovery                                  node1
dsconfigurator                             node1
monitor                                    node1
resource.node1                             node1
scheduler                                  node1
searchapp.node1                            node1
searchmanager.node1                        node1
searchserver.node1                         node1
utilities.node1                            node1
FFQC5324I  -------- End of Report. Total: 23 --------
```

If you defined multiple crawlers of the same type, for example, you define three Windows file system crawlers in one collection, you still see several session IDs for those crawlers. In such case, you can see the display name of the crawler session ID by using the following command:

```
esadmin report sessions -sid session ID -format full
```

Example 15-8 shows the output of using this command.

Example 15-8 Result of the esadmin report sessions command with -format full option

```
>esadmin report sessions -sid col_sample.WIN_77148 -format full
FFQC5323I Sessions Report

-----------------------------------------------
Session Id:    col_sample.WIN_77148
Display name:  IVPFile System
Description:
Collection Id: col_sample
Node Id:       node1
Type:          crawler
Subtype:       WIN
User:
Password:
Domain:        .
Flags:         0
Config dir:    col_sample.WIN_77148
Data dir:
Log dir:
Target:
Properties:    {init_heap=16, max_heap=512}

FFQC5324I  -------- End of Report. Total: 1 --------
```

Alternatively, when you want to see all details of the crawlers regardless of the collection, you can also enter the following command:

```
esadmin report sessions -type crawler -format full
```

For further details about the **`esadmin report`** command, type the following command at the command prompt:

```
esadmin report help
```

15.7.3 General guidelines for monitoring the operating system

This section provides general guidelines for monitoring the operating system. Depending on the situation you run into, you might require much deeper system monitoring and analysis with help from a performance analyst. Also, because we do not cover the details of each command, see the product documentation for further information about each command.

System monitoring on AIX or Linux

Generally, when you use Content Analytics on AIX or Linux, the **vmstat** command and **iostat** commands are helpful to monitor the entire system resource usage. For example, they can monitor processor utilization, memory usage, enough I/O throughput, and so on. With these commands, you can confirm whether the system has enough memory resources.

Also, the **ps** command has various options to help identify how each process uses the system resource, because the **vmstat** command or **iostat** command do not indicate which process is using resources.

For example, consider the following situations:

- You use the **vmstat** command to confirm that 80% of processor resources are used already. This processor utilization is extremely high compared to the usual state. Therefore, you want to see which process uses the resource. In such case, you can use the **ps auxww** command or **ps -elf** command to help you identify which process uses the system resources, such as memory or processor. You can sort the results by the resource that you want to see. After you identify a process that uses the resource a lot, consider why the usage occurs.
- You use the **vmstat** command to confirm that 50% of processor resources are used, which is not as high of utilization as you expect. Therefore, you check the **iostat** output and confirm that some disk I/O is taking a long time. You investigate why the disk I/O is taking so long and block the processor utilization.

System monitoring on Windows

When you use Content Analytics on Windows and then monitor the system resources, start with the task manager or the **tasklist** command.

If you need further analysis for the system resources in detail, use the **perfmon** utility to help you monitor system resource utilization and the detailed resource usage per process.

For AIX or Linux cases, proceed with the investigation the same as described for Windows. However, you might have to set up the **perfmon** utility to fit your purposes. Various performance objects and counters are associated with the objects. Therefore, you must select the suitable performance objects and counters.

16

Hints and tips for troubleshooting

This chapter provides hints and tips that you can use when troubleshooting problems that you might encounter during IBM Content Analytics administration and usage.

This chapter includes the following sections:

- Overview of troubleshooting
- General troubleshooting guidelines
- Working with the logs in Content Analytics
- Installation and administration-related troubleshooting
- Text miner application-related troubleshooting
- Data processing flow-related troubleshooting
- Export-related troubleshooting
- Classification Module server-related troubleshooting tips
- Reporting a problem to the IBM Software Support
- Advanced troubleshooting topics

16.1 Overview of troubleshooting

To help you quickly troubleshoot your problems, we list the questions and problems that you might encounter when working with Content Analytics. Look through the list to see if any of them is applicable to you. Then go to the specific page for the problem description and answer to your problem.

> **Tip:** To successfully troubleshoot your problems, before you proceed with the troubleshooting task, see 16.2, “General troubleshooting guidelines” on page 605, and 16.3, “Working with the logs in Content Analytics” on page 607.

- Installation and administration-related troubleshooting
  - “How do I install Content Analytics with a larger temporary disk space?” on page 610
  - “Why can’t I start Content Analytics on a single server?” on page 611
  - “Why doesn’t Content Analytics start on the additional server?” on page 613
  - “Why can’t I create a crawler such as for DB2 or Notes?” on page 614
  - “Why can’t the session be started?” on page 614
- Text miner application-related troubleshooting
  - “Why can’t I start Content Analytics on my browser?” on page 615
  - “Why can’t I see facets in the text miner application?” on page 616
  - “How do I get detail messages of the text miner application?” on page 617
  - “What do I do when I see the ‘Unknown error occurred’ message?” on page 617
  - “What do I do when I see an error message from the browser?” on page 618
- Data processing flow-related troubleshooting
  - “Why do I see the TYPE_IO_ERROR message during crawling?” on page 619
  - “How do I verify if a specific document is crawled?” on page 620
  - “Why doesn’t the crawler crawl the document in question?” on page 621
  - “Why does the resource deployment fails?” on page 622

  - "Why are the documents dropped in the parse and index process?" on page 623
  - "How do I confirm if a specific document is indexed?" on page 624
- Export-related troubleshooting
  - "Why does the parse and index component frequently fail to start after configuring a custom export plug-in?" on page 624
  - "Why can't I export documents after a configuration changes?" on page 625
- Classification Module server-related troubleshooting tips
  - "Verify that the decision plan is up and running" on page 626
  - "Verify that the associated knowledge bases are running" on page 626
  - "Use trace to diagnose Classification Module server issues" on page 626

If the information does not help solve your problem, see 16.7, "Export-related troubleshooting" on page 624, and 16.10, "Advanced troubleshooting topics" on page 628, for additional information.

16.2 General troubleshooting guidelines

When you troubleshoot a problem, follow these guidelines:

- Clearly understand the situation when the problem occurs.

  Gather information, such as the operation you perform, the message you see, or when exactly the problem occurs. If you do not see any messages, you must be clear in regard to what you see on the panel, especially when you use the text miner application. Clearly understand the environment such as your Content Analytics version, the data source type, and the operating system platform.

- Confirm the log files to see if any messages are logged.

  If you see any guidance in the message, follow the suggested step to see if you can resolve the problem.

  Optionally, if you identify the component in question, you might enable the extra traces to collect additional information and then reproduce the problem. With the traces enabled, you might see more detailed messages in the log or the trace, which helps you to resolve the problem or take additional action.

- If the message does not include a suggestion, search the web or knowledge bases to see if you can find any solution for the corresponding messages.

You might want to see if it is a known issue as explained in the *Problems and solutions in IBM Content Analytics, Version 2.2* technote at following web address:

http://www-01.ibm.com/support/docview.wss?uid=swg27017955

If you cannot find the solution to the problem by yourself, contact IBM Software Support.

These guidelines are general. Depending on the problem, the troubleshooting approach can be different. This section describes the points that help you to understand the problem and environment in general.

16.2.1 Understanding the problem

When you perform troubleshooting, you must understand the problem clearly and the details of the problem. Knowing the following details can help you explain the problem:

- If you see a particular message, the precise message, for example, FFQR0160E
- The operation that you performed at that time
- When you see the message
- Whether you have seen the same message before or this time is the first time you have seen the message:
  - If this time is the first time you that you have seen the message, confirm whether you changed any part of the configuration before the operation.
  - If this message occurs frequently, see how frequently does it occur, such as every time you perform the same operation, once a month, and so on.
- Whether the problem reproducible

  That is, the problem occurs every time you perform the same operation. Make sure you can reproduce the problem.

Window capture: Sometimes the problem is related to the text miner application window or the administration console. In either case, taking a window capture is helpful to explain the problem to IBM Software Support.

16.2.2 Understanding the environment

When you troubleshoot a problem, consider your environment, such as the Content Analytics version, fix-pack level, server type (single server installation or distributed server installation), and operating system environment. Understand where the related components are located and narrow down the problem area.

The environment information of the data source is sometimes important for understanding the problem overview. When you crawl the various data sources, the type of the data source is important if the problem is related to a particular data source.

Collect the following environment-related information:

- Operating system type and version, including patches and service pack levels
- Content Analytics version, fix pack level, and applied fix packs
- Server configuration (single server or multiple servers)
- Data sources in your collections, including the repository or data source type, version, and operating system
- The number of collections defined in your Content Analytics system
- The collection type (search collection or text analytics collection)

  When you use the text miner application, you must have at least one text analytics collection at least.
- The collection ID of the collection where the problem occurs

16.3 Working with the logs in Content Analytics

Usually if a problem occurs with a message, the component writes an entry in a log file. See if any related messages are logged with the time line of when you see the problem.

This section describes the following log-related topics:

- The location of the logs
- Understanding the log contents
- The first log to examine

16.3.1 The location of the logs

During the operation, Content Analytics-related log files are in the `ES_NODE_ROOT/logs` directory. If the log is a centralized log, such as a system log or collection log, it is in the `ES_NODE_ROOT/logs` directory on the master server.

For some session component-related logs, the logs are in the `ES_NODE_ROOT/logs/audit` directory. These log files are session-dependent and are on each server.

Table 16-1 shows the location of the major log files. The variable *yyyymmdd* represents the date on which the log file is created.

Table 16-1 Log file name, location, and explanation

Log file name	Log location	Description

`system_`*yyyymmdd*`.log`
Example:
`system_20100610.log`
	`ES_NODE_ROOT/logs` directory on the master server	A system log that records system-wide events.

collectionID_yyyymmdd`.log` Example: `col_sample_20100610.log`	`ES_NODE_ROOT/logs` directory on the master server	A collection log that records the collection-related session events.

`ccl_`*hostname_N*`.log`
Example:
`ccl_cca.example.com_0.log`
	`ES_NODE_ROOT/logs` directory on the master server	A common communication layer (CCL) log that records the CCL-related events.

sessionID`_audit_`*yyyymmdd*`.log` Example: `col_sample.WIN_77148_audit_20100610.log`	`ES_NODE_ROOT/logs/audit` directory on each server	A session log that records each session-related event. After you identify the session in question, check the entries.

You might also want to check the following logs depending on the problem:

- Installation-related log files in the `ES_NODE_ROOT/logs/install` directory on each server, which help you to understand what occurs during the installation
- The text miner application or search application-related logs, such as the `search*` log files, in the `ES_NODE_ROOT/logs/jetty` directory on the search server
- The administration console-related logs, such as the `admin*` log files, in the `ES_NODE_ROOT/logs/jetty` directory on the master server

16.3.2 Understanding the log contents

In Content Analytics, some logs are still human readable, while others are not. If the logs are not human readable, use the utilities to understand the log content.

You can view the log content in the following ways, depending on the type of the log file:

- View the logs from the administration console on the master server.
- View the logs with the commands on each server.

Viewing the log files from the administration console

You can view log content from the administration console if the log is the system log or the collection log. These logs are the starting point when you check the log files. For more information, go to the IBM Content Analytics Information Center at the following address, and search on *viewing log files*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Viewing the log files by using commands

You can only see the system log and collection log from the administration console. For some log files, you can view their messages from the command line by using the **esviewlog** command or the **esviewauditlog** command:

- For AIX and Linux platforms, use the **esviewlog.sh** and **esviewauditlog.sh** commands.
- For the Windows platform, use the **esviewauditlog.bat** command.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *viewing log files without using the administration console*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

16.3.3 The first log to examine

You now understand where the logs are located and how to view their content, but you see many files. Where do you start?

In general, first check the system log and the collection log from the administration console, or use the **esviewlog** command. You can find the system log and the collection log in the `ES_NODE_ROOT/logs` directory on the master server. Make sure that you view the log file of the day when the problem occurs. The system log and collection log record the session ID. You can check the component log after you identify the collectionID or sessionID.

You can find the session log in the `ES_NODE_ROOT/logs/audit` directory on each server. You must format these logs by using the **`esviewauditlog`** command to see if any messages are logged. Make sure that you view the session log file of the day when the problem occurs. If the message suggests any actions, follow the suggestion and see if the problem is resolved.

Review the formatted log, and at a minimum, find the time frame in the log entries when the problem occurs.

> **Maximum log file size:** By default, the maximum log file size is 16 MB, and the maximum number of log files is 10. (Up to 10 files are created for a particular log.) When the file size reaches the maximum log file size, the old log file is renamed *`filename.N`*, and a new file is created.
>
> The oldest log is removed when the number of files reaches the maximum. To change the maximum log file size or the maximum number of files, see 16.10, “Advanced troubleshooting topics” on page 628.

16.4 Installation and administration-related troubleshooting

This section provides practical troubleshooting tips related to system installation and administration, such as starting the system and configuring a data source crawler. This section addresses the following questions:

- How do I install Content Analytics with a larger temporary disk space?
- Why can’t I start Content Analytics on a single server?
- Why doesn’t Content Analytics start on the additional server?
- Why can’t I create a crawler such as for DB2 or Notes?
- Why can’t the session be started?

How do I install Content Analytics with a larger temporary disk space?

Question: When I tried to install Content Analytics with less temporary disk space, it failed. How do I install Content Analytics to have a much larger temporary disk space?

Answer: Set the `IATEMPDIR` environment variable to point to a different temporary disk space, and invoke the installer.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *running out of space during installation*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Why can't I start Content Analytics on a single server?

Question: When I try to start the system using the **esadmin system startall** command on a single server system, I see the FFQC5363E message (Example 16-1). Content Analytics worked for a while after I installed it.

Example 16-1 Text message when starting Content Analytics fails

```
C:\Documents and Settings\Administrator>esadmin system startall
FFQC5302I Starting the system...
FFQC5362I Starting the Common Communication Layer (CCL) on
cca.imecm.local. Wait for the command that is running [cmd /c call
startccl.bat] to complete on cca.imecm.local.
FFQC5363E The request on cca.imecm.local to start the Common
Communication Layer (CCL) failed. More details are available in
C:\IBM\es\esadmin\logs\trace\startstatus.log.
null
FFQC5391E The information center on the search server cannot be
accessed. Start the Common Communication Layer (CCL) service on
cca.imecm.local.
FFQC5395E One or more of the required system services are not running.
Review the error and warning messages to see which services are not
started and to learnmore about the problem. For additional information,
see the log file C:\IBM\es\esadmin\logs\trace\startstatus.log.
```

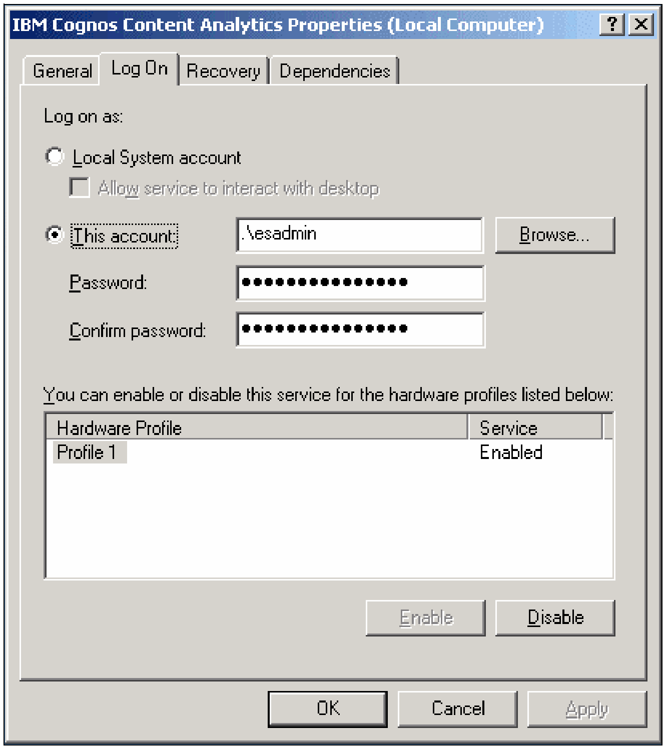
The message indicates that one of the components, in this case, the CCL component, cannot start with the command. What should I do to start Content Analytics with the command?

Answer: Typically you cannot start Content Analytics because the password of the administrative user expired. Verify whether you can log in as the administrative user with the password that you set before. If you use Content Analytics on the Windows platform, check the Event viewer to confirm what happens.

When you change the password of the administrative user at that time, set the new password by using the **eschangepw** command (**eschangepw.sh** for AIX and Linux, and **eschangepw.bat** for Windows).

Also, when you use Content Analytics on the Windows platform, change the password. Make sure that you change the password on the control panel. To change the password, follow these steps:

1. Select **Start** → **Administrative Tools** → **Services**.
2. Right-click **IBM Content Analytics** and select **Properties**.
3. In the Properties window (Figure 16-1), select the **Log On** tab, and change the password settings. Then click **OK**.



IBM Cognos Content Analytics Properties (Local Computer)

General | Log On | Recovery | Dependencies

Log on as:

○ Local System account

□ Allow service to interact with desktop

● This account: .\esadmin [Browse...]

Password: ●●●●●●●●●●●●●●●

Confirm password: ●●●●●●●●●●●●●●●

You can enable or disable this service for the hardware profiles listed below:

Hardware Profile	Service
Profile 1	Enabled

Figure 16-1 Password settings in the Properties window

For details about using the **eschangepw** command, go to the IBM Content Analytics Information Center, and search on the following topics depending on your environment:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

- "Changing the administrator password in a single server configuration"
- "Changing the administrator password in a multiple server configuration"

When you install Content Analytics on Windows 2008 server, log in as the administrative user (`esadmin` by default) at least once before you start Content Analytics.

Often, users forget to log in as the administrative user and try to start Content Analytics as the Administrator user of the operating system that they used to install the product. However, you must log in first as the administrative user that you set during the Content Analytics installation.

For the installation steps, go to the IBM Content Analytics Information Center at the following address, and search on *installing the product on a single server*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Why doesn’t Content Analytics start on the additional server?

Question: When I try to start the system using the **`esadmin system startall`** command on the master server, I never succeed in starting the system on the additional server, while the command starts the sessions on the master server. What is the possible cause for this problem, and what can I do to correct it?

Answer: When you install an additional server, you must specify the same user ID and password that you used during the master server installation. To solve the problem, consider reinstalling the additional server with the correct user ID and password.

Alternatively, if you use the same user ID but a different password on the additional server, you can temporarily change the password on the additional server to see if you can start Content Analytics on the additional server.

> **The CK property value:** The CK property value is determined when you install the system, and it varies depending on the system. You must use the same user ID and password when you first install the additional server.
>
> **Tip:** Although you can quickly use the following procedure to fix the problem, the solution is for test purposes only. To solve your problem permanently, reinstall the additional server with the correct user ID and password if the additional server is for production purposes.

To see if you can start the additional server, follow these steps:

1. Stop CCL on the additional server by using the **`stopccl`** command. Alternatively, stop CCL from the Windows services panel if the additional server is on Windows.
2. Confirm that the CK property, such as `CK=`*`String`* in the `ES_INSTALL_ROOT/nodeinfo/es.cfg` file, is on the master server.

3. Open the `ES_INSTALL_ROOT/nodeinfo/es.cfg` file on the additional server, and copy the `CK` property value from the master server.
4. Enter the **eschangepw** command with the correct password (that is used on the master server) on the additional server.
5. Restart CCL on the additional server by using the **startccl** command. Alternatively, start CCL from the Windows services panel if the additional server is on Windows. See if you can also start the system on the additional server.

Why can't I create a crawler such as for DB2 or Notes?

Question: When I try to create a DB2 crawler from the administration console, I get the FFQD2021E message on the administration console (Example 16-2).

Example 16-2 FFQD2012E message when creating DB2 crawler

```
FFQD2012E The crawler server is not configured for the specified
crawler type. Ensure that the crawler server is configured by the DB2
setup script: escrdb2.vbs (Windows) or escrdb2.sh (Linux and AIX).
```

Why can't I create a crawler, and what should I do?

Answer: In this case, the message is quite descriptive. Run the necessary script for your data sources, such as the **escrdb2.*** script for DB2 or the **escrnotes.*** script for Lotus Notes (especially when you use Notes Remote Procedure Call (NRPC)). Run the script depending on the data source type before you configure the crawler.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *crawler administration*:

http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Why can't the session be started?

Question: When I try to start the sessions with the **esadmin system start** command, it fails. Before starting the sessions, one of the sessions was abnormally terminated (due to a crash). How do I start the session in such a case?

Answer: When the session is abnormally terminated, you must run the cleanup procedure before you restart the session.

To clean up the session, issue the **esadmin *sessionID* destroy** command against the session in question. Then issue the session starting command, such as the **esadmin system start** command to start all sessions or the **esadmin *sessionID* start** command to start a specific session again.

16.5 Text miner application-related troubleshooting

This section provides common troubleshooting tips when you encounter a problem while using the text miner application:

- Why can’t I start Content Analytics on my browser?
- Why can’t I see facets in the text miner application?
- How do I get detail messages of the text miner application?
- What do I do when I see the ‘Unknown error occurred’ message?
- What do I do when I see an error message from the browser?

For more information about the text miner application, see Chapter 5, “Text miner application: Basic features” on page 143.

Why can’t I start Content Analytics on my browser?

Question: When I try to start the text miner application, the browser freezes while loading the page. Even after the text miner application window opens, I cannot see the Text Analytics views such as the Facets view. What can I do in such cases?

Answer: Check the browser version and the prerequisite software. Confirm the following points:

- Check the browser version that you use and see if it is listed as the supported version.

  You can see the supported browser from the *IBM Content Analytics Version 2.2 system requirements* page at the following web address:

  http://www-01.ibm.com/support/docview.wss?rs=4173&uid=swg27017944

  The supported browser is updated from time to time. Therefore, the latest version of a browser or an older version might not be supported. Check the page to confirm the supported browsers. At GA time of Content Analytics Version 2.2, Microsoft Internet Explorer Version 7 and 8 and FireFox Version 3.5 and 3.6 are the supported browsers.

- Check the Adobe Flash Player version that you use on the browser, and see if it is Flash Player 10 or later.

  You can check the Adobe Flash Player version at the following web page:

  http://kb2.adobe.com/cps/155/tn_15507.html

  If you are not using Adobe Flash Player 10, update Adobe Flash Player.

- Verify that JavaScript support is enabled in a browser setting.
  - If you use Microsoft Internet Explorer:
    - Determine the zone, such as the Internet or intranet, to which Content Analytics belongs.
    - Verify if JavaScript is enabled in the zone. JavaScript is not enabled if the Content Analytics page belongs to a zone that has a security level of High.
    - If you use Microsoft Internet Explorer on the Windows Server operating system, such as Windows 2003 Server or Windows 2008 Server, the default security level for the Internet zone is set as High. Add Content Analytics pages to *Trusted Sites* or *Local Intranet* if the zone is the Internet.
  - If you use FireFox, select **Tools** → **Options**. Select the **Content** tab and verify whether the **Enable JavaScript** check box is selected.

Why can't I see facets in the text miner application?

Question: I defined facets and performed crawl, parse, and index. Then I did a search for the text analytics collection. However, when I access the text miner application, I cannot see the facets that I defined. What do I do in such a situation?

Answer: Sometimes, especially when you define many facets, more time is required to load all the facets. Wait for a while until all facets are loaded into the application.

You can see the progress of uploading the facets in the `ES_NODE_ROOT/logs/audit/`*`collectionID`*`.searcher.node`*`N`*`_audit_`*`yyyymmdd`*`.log` file. Search for keywords, such as those in Example 16-3, in the file after you format it by using the **`esviewauditlog`** command.

Example 16-3 The audit log entry that shows the progress of Facet uploading

```
FFQX0717I The doc freqs is being loaded for the facet counter
FFQX0717I Calculation of doc freqs has been completed
```

How do I get detail messages of the text miner application?

Question: If I want to see the detailed messages of the text miner application, how do I see them? Also which log do I check?

Answer: You can see the logs in the `ES_NODE_ROOT/logs/jetty` directory on the search server. The following logs help you to see the detailed messages:

- `searchapp.node1.searchapp.0.log`
- `searchapp.node1.textminer.0.log`
- `searchserver.node1.ESSearchServer.0.log`

> **Log location:** If you deploy the text miner application on WebSphere Application Server, confirm that the WebSphere Application Server log, such as the `SystemOut.log` file or `SystemErr.log` file, is in the *`WAS_PROFILE_ROOT`*`/logs/ESSearchServer` directory depending on your environment.

What do I do when I see the ‘Unknown error occurred’ message?

Question: When I use the text miner application, I receive the “`Unknown error occurred`” message, and I cannot proceed with the analysis. What do I do when I encounter such a situation?

Answer: The message is displayed because some sessions working for the text miner application disappeared for some reason. In most cases, this problem occurs because of an issue with the Java virtual machine (JVM). You might want to turn off the Just-In-Time (JIT) compiler of the session to work around the problem.

To turn off JIT compiler of the session, modify the configuration file as follows:

1. Go to the `ES_INSTALL_ROOT/configurations/interfaces` directory on the search server, and find the `searchapp__interface.ini` file and the `searchserver__interface.ini` file.
2. Open each file, and add `JVMOptions=-Xint` at the end of each the file. Then save the changes to each file.
3. Restart the `searchapp.node`*`N`* session and `searchserver.node`*`N`* session. See whether you encounter the same problem. You can restart the session by using the **`esadmin`** ***`sessionID`*** **`stop`** and **`esadmin`** ***`sessionID`*** **`start`** commands as shown in Example 16-4.

   Example 16-4 Restarting the session with the esadmin command

```
>esadmin searchapp.node1 stop
FFQC5303I Search Application (node1) (sid: searchapp.node1) CCL
session exists.
```

```
PID: 1388
FFQC5314I The following result occurred: 0
>esadmin searchapp.node1 start
FFQC5310I Search Application (node1) (sid: searchapp.node1) is not
running.
FFQC5314I The following result occurred: 0
```

If the problem persists after you disable the JIT compiler of the session, contact IBM Software Support for further assistance.

Same problem in administration console: If you encounter the same problem when using the administration console, try to disable the JIT compiler for administration session. That is, modify the `ES_INSTALL_ROOT/configurations/interfaces/admin__interface.ini` file on the master server to disable the JIT compiler. Then restart the session.

What do I do when I see an error message from the browser?

Question: When I use the text miner application, I receive the "`Syntax error`" message. This message is in a smaller pop-up window and is a different message than discussed in the previous section. What do I do when I encounter such a situation?

Answer: The message indicates that the error occurs at the browser side, not the server side. Typically in this case, no log entries are in the log files at the server side. You must check the error console of the browser and determine how to reproduce the problem. After you determine the following information, contact IBM Software Support for further assistance.

- Does the problem occur on a specific supported browser? For example, if you experience the problem with FireFox, confirm whether you see the same error with Microsoft Internet Explorer.
- If the problem occurs with FireFox, open the error console output as follows:

  a. Reproduce the problem on the FireFox.
  b. Select **Tools** → **Error Console**, and select the **Errors** tab.
  c. Save the error text in the **Errors** tab.

16.6 Data processing flow-related troubleshooting

Sometimes you cannot find a document in the Documents view in the text miner application even though you know that the document exists. If you encounter such a search result-related problem during your analysis, what must you do?

In that case, identify where the problem occurs in the data flow. For example, to analyze the document in the text miner application, you must index the document. To index the document, you must crawl the document first, then pass it through the parse and index process, and finally complete parsing and indexing successfully. Check if the document is crawled. Then consider if the document is parsed and can finally be indexed.

This section describes the troubleshooting tips from the data flow perspective:

- Why do I see the TYPE_IO_ERROR message during crawling?
- How do I verify if a specific document is crawled?
- Why doesn't the crawler crawl the document in question?
- Why does the resource deployment fails?
- Why are the documents dropped in the parse and index process?
- How do I confirm if a specific document is indexed?

Why do I see the TYPE_IO_ERROR message during crawling?

Question: While the crawler is running, I see a `TYPE_IO_ERROR` with `SiapiException`, and I see the `FFQM5035E` message from the crawler when I check the collection log from the administration console. What do I do when I see this message?

Answer: This error message can occur when the temporary storage for the crawled data is full. Make sure that the parse and index process is running. If the parse and index process is already running and you see the message, increase the temporary storage for the crawled data.

You can change the upper limit from the administration console as follows:

1. Log in to the administration console, and select the **System** link on the top.
2. Select the **DataListener** tab with the edit mode, and click the **Configure Data Listener applications** link.
3. Set the "Maximum amount of data in temporary storage, per collection value" field to a larger value. By default, `10000 MB` (10 GB) is set. Depending on your system environment, increase the value.
4. Restart the DataListener session by using the **`esadmin datalistener stop`** and **`start`** commands, or by using the **`esadmin system stop`** and **`start`** commands to restart all sessions.

Temporary storage size: When you define many facets for your analysis and crawl a large amount of data, you might see the `TYPE_IO_ERROR` message with the default temporary storage size. Perform the previous steps before you start crawling.

How do I verify if a specific document is crawled?

Question: How can I verify if a specific document is crawled by the crawler that is configured to crawl the document?

Answer: Suppose that you know the file name in question. You can confirm the URL of the crawled document by using the **`esadmin rds read`** command as explained in the following steps:

1. Stop all crawlers defined in the collection.
2. Run the parse and index process to complete current data processing.
3. After the previous step is completed, stop the parse and index process.

> **Important:** You must stop the parse and index process before you start the crawl process in step 4. Otherwise, all crawled data is processed right away.

4. Start the crawler session in question, and perform a full crawling.
5. Stop the crawler session after the full crawling is completed.
6. Run the following command:

   ```
   esadmin rds read -cid collectionID -url -httpcode
   ```

   Example 16-5 shows the **`esadmin rds read`** command running against the crawler data of the Windows file system. Verify in the result whether the URL of the target document is found.

   Example 16-5 Output of the esadmin rds read command

   ```
   >esadmin rds read -cid col_sample -url -httpcode
   URL:
   file:///C:/IBM/es/samples/firststep/data/xml/xmls.tar.gz?ArchiveEntr
   y=xml%2F00000000.xml
   HTTP Code: 2000
   URL:
   file:///C:/IBM/es/samples/firststep/data/xml/xmls.tar.gz?ArchiveEntr
   y=xml%2F00000001.xml
   HTTP Code: 2000

   .....

   URL:
   file:///C:/IBM/es/samples/firststep/data/xml/xmls.tar.gz?ArchiveEntr
   y=xml%2F00000850.xml
   HTTP Code: 2000
   ```

```
URL:
file:///C:/IBM/es/samples/firststep/data/xml/xmls.tar.gz?ArchiveEntr
y=xml%2F00000851.xml
HTTP Code: 2000
Found 852 document(s) total
```

7. After you confirm the URL in the **`esadmin rds read`** command output, start the parse and index process again to proceed with the data flow.

If you do not find the URL in question in the command result, the crawler does not crawl the document in question. You need to pursue why it occurs.

Why doesn't the crawler crawl the document in question?

Question: When I start the crawler and confirm the crawled data with the **`esadmin rds read`** command, why can't I find the document URL in question?

Answer: When you use the data source crawler (other than the web crawler), consider following points:

- Check whether the target document is included in the crawl space.

  Confirm if the target document in question is in the crawl space. The crawl space definition depends on the data source crawler. For example, the crawl space for the Windows file system crawler is a *folder*. Therefore, you must verify if the document is in the folder or if the archive file is in the crawl space.

  Some crawlers, such as the seed list crawler, WebSphere Portal crawler, or WebSphere Content Manager crawler, use the seedlist to get the document list to crawl. In such case, confirm if the target document is included in the seedlist.

- Check which crawl mode you use for crawling.

  The crawler starts the crawling in the all-update-crawling mode (for all new documents, modified documents, and deleted documents) when the crawler session is started. If the document in question is not updated since the last crawling, the document is not crawled in the all-update-crawling mode. In such case, try a full crawling and see if the document in question is crawled.

  If the document in question is deleted, but you start the crawler with the new-and-modified-update-crawling mode, the deleted document is not detected.

  In conclusion, when you crawl the documents with the data source crawler, know which crawl mode you use. For more information about the crawling mode and schedule, go to the IBM Content Analytics Information Center at the following web address and search on *crawler schedules*:

  http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp

Perhaps the document is found in the crawl space and you performed a full crawling, but the document is still not crawled. In this case, confirm whether the crawler user has enough privileges to connect to the data source and extract the data from the data source. The crawler user is the user that is specified in the data source crawler configuration.

When you use the web crawler, troubleshooting is different. The web crawler continues crawling while the crawler session is up and running, and the crawling schedule strategy is different from the data source crawler. The next crawling schedule of the web page is determined based on the recrawl interval settings in the web crawler.

If you do not want to wait until the next crawling schedule, select the **URLs to visit or revisit** option after you start the web crawler.

For more information, go to the IBM Content Analytics Information Center at the following address, and search on *options for visiting URLs with the web crawler*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

Why does the resource deployment fails?

Question: When I run the Resource deployment, it fails with following message:

```
FFQI0063E Cannot write back resource files to the master configuration
directory
```

What is the possible cause of the failure, and what do I do?

Answer: When a process (such as an editor or shell) locks the `ES_NODE_ROOT/master_config/`*`collectionID`*`.indexservice/resource` directory or its subdirectories, you see the `FFQI0063E` message, and the resource deployment fails.

Make sure that other processes do not use the `ES_NODE_ROOT/master_config/`*`collectionID`*`.indexservice` directory or its subdirectory. Consider the following typical cases:

- You open a command prompt or shell on the directory, or edit files in the directory.
- You use Content Analytics on Windows, and open Windows Explorer in the directory to see files.

Why are the documents dropped in the parse and index process?

Question: On the Parse and Index Details page on the administration console, I notice that a value for the “Number of dropped documents” field in the Parse and index status summary section is not 0. This value means that some documents are dropped in the parse and index process. Why are the documents dropped in the parse and index process? What is the possible cause?

Answer: The document can be dropped in the parse and index phase for several reasons. For initial troubleshooting, confirm the following points:

- The document type is supported in Content Analytics.

  Verify if the document type is listed in IBM Content Analytics Information Center at the following address by searching on *default support document types*:

  `http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

- The document is not corrupted.

  Confirm whether you can see the document in the data source without any problem. For example, confirm if you can open and see the PDF file with an application such as Adobe Reader.

- The document is not protected by the password.

  For example, you might want to set a password for the Microsoft Excel document. However, this kind of protected document can be dropped.

- The document is well formatted and can be parsed, especially when you crawl the XML files.

  If the document is not well formatted and cannot be parsed by XML parser, the document is dropped.

If the document in question does not apply to any of these cases, sometimes the document is dropped still. Consider the example where the document is not corrupted and you can open the file with a suitable application or the document is well formatted. In such case, contact IBM Software Support for further assistance.

How do I confirm if a specific document is indexed?

Question: How can I confirm if a specific document is indexed?

Answer: Launch the text miner application. Go to the Documents view, and search with keyword `docid:`*`documentID`*. *`DocumentID`* in the URL as you see in the output of the **`esadmin rds read`** command.

For example, if you want to verify if the `00000500.xml` file is found in the index, search with following query to see if the file is in the search result:

```
docid:file:///C:/IBM/es/samples/firststep/data/xml/xmls.tar.gz?ArchiveE
ntry=xml%2F00000500.xml
```

You can also use other information, such as title field, if you are unsure about the document ID for the target document in question.

If you cannot find the document in the search result, verify that the document is processed (that is, crawled, parsed, and indexed) as explained earlier.

16.7 Export-related troubleshooting

This section describes common problems that you might come across when using the export feature of Content Analytics. This section addresses the following questions:

- Why does the parse and index component frequently fail to start after configuring a custom export plug-in?
- Why can't I export documents after a configuration changes?

We provide guidance for troubleshooting these problems.

Why does the parse and index component frequently fail to start after configuring a custom export plug-in?

Question: After configuring a custom export plug-in, why does starting the parse and index component fail frequently? The following exception is displayed in the system log:

```
FFQO0277E An exception was caught with the detail
'com.ibm.es.oze.indexservice.internal.IndexServiceImpl$IndexBuildCouldN
otStartException: FFQEI0006E' and a stack trace of
'com.ibm.es.oze.indexservice.internal.IndexServiceImpl$IndexBuildCouldN
otStartException: FFQEI0006E at
com.ibm.es.oze.indexservice.internal.IndexServiceImpl.startIndexBuild(I
ndexServiceImpl.java:892) at
```

```
com.ibm.es.control.indexservice.server.ComponentIndexService.startIndex
Build(ComponentIndexService.java:293) at
com.ibm.es.control.indexservice.server.ComponentIndexServiceW.startInde
xBuild(ComponentIndexServiceW.java:47) at
.......
```

Answer: This issue occurs when the custom plug-in incorrectly handles information about deleted documents. A custom export plug-in performs processing on the document object that is being exported from Content Analytics. If the crawler is configured to crawl new, modified, and deleted documents, Content Analytics also passes a null object for the deleted documents. In this case, the implementation of the custom export plug-in must handle the null document case. Use the following statement to check whether it is a normal document or deleted document:

```
com.ibm.es.oze.api.export.document.Document#getType()
```

The other option is to disable the exporting information about deleted information. You can do this in two ways: configuring the crawler to only crawl new and modified data or configuring the export options to not include information about deleted documents. See 10.2.1, “Crawled documents” on page 397, which explains crawler configuration. Also see 10.4.4, “Exporting information about deleted documents” on page 406, which explains options for not exporting information about deleted documents.

Why can’t I export documents after a configuration changes?

Question: After configuring or changing the export option in the collection, why are documents not being exported even though the crawler, parse and index, and search components are running?

Answer: If you configure or change the export options for crawled or analyzed documents *after* the collection is built, restart the parser and index component of the collection for the changes to take effect. You do not need to restart the parse and index component if you make changes to export options for search results.

16.8 Classification Module server-related troubleshooting tips

This section provides troubleshooting guidance for using the Classification Module server.

Verify that the decision plan is up and running

Make sure the decision plan that you intend to use with the Content Analytics collection is up and running. Launch the Classification Module Management Console to make sure that the system connects to the Classification Module Server URL that you configured in the Content Analytics. When Classification Module Management Console starts, the connection configuration is displayed. Check that the *server name* and *port* are the same as you configured inside the Content Analytics collection that you work with.

Verify that the associated knowledge bases are running

Make sure that both the decision plan and its associated knowledge bases are running.

From the Classification Module Management Console, select the **Decision plans** icon in the left panel to obtain information about all the decision plans deployed on the Classification Module server that you connected to. Make sure the decision plan that you work with is listed and its status is *Started*.

From the Classification Module Management Console, select the **Knowledge bases** icon in the left panel. Notice the list of associated knowledge bases for the decision plan you are working with. Verify that all of these knowledge bases are running.

For more information about usage, go to the IBM InfoSphere Classification Module Information Center at the following address and search on *Management Console*:

http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp?topic=/com.ibm.classify.admin.doc/c_AG_classification_manager.htm

Use trace to diagnose Classification Module server issues

You can engage the Classification Module Trace application to diagnose problems in a production environment. For more information, see *Collecting data for InfoSphere Classification Module* at the following web address:

http://www.ibm.com/support/docview.wss?uid=swg21417244

16.9 Reporting a problem to the IBM Software Support

When you cannot find the solution for the problem, you can report the problem to IBM Software Support. Prepare the following information before you contact the support team so that the support representative can understand and solve your problem more efficiently:

- Prepare a detailed problem description as explained in 16.2, “General troubleshooting guidelines” on page 605.
- Run the **`esservice`** command with the **`-nocore -noheapdump`** option on the master server and on each server where the problem occurs when you use the distributed server configuration. Collect the compressed output file (`.zip` file).

  The **`esservice`** command utility is available as the **`esservice.sh`** command for the AIX and Linux platforms and as **`esservice.bat`** for the Windows platform.

  When you do not specify the **`-file`** option, the **`esservice`** command creates the `service_`*`component_yyyyMMdd_HHmmssz`*`.zip` file in the directory where you run the command. The *`component`* value represents the type of node, such as `controller`, `crawler`, or `search` in the distributed server configuration. If you use multiple server configurations, such as having two search servers, make sure that you know which server you collect the **`esservice`** command output from.

  The file size of the `output.zip` file tends to be bigger if many log files or the large files are included in the `ES_NODE_ROOT/logs` directory. Make sure that you have enough space to save the output.zip file.

You might need to provide additional information or trace depending on the nature of the problem.

For more information about the **`esservice`** command, go to the IBM Content Analytics Information Center at the following web address, and search on *gathering information for problem analysis*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

16.10 Advanced troubleshooting topics

This section explains the advanced topics for troubleshooting in Content Analytics. To begin, check the existing logs to see if you can find any indication for the problem. If the existing logs do not provide enough troubleshooting information, collect additional detailed logs when reproducing the problem.

The following techniques are described for collecting detailed information:

- Changing the log level of the system and collection logs
- Generating a javacore or a heapdump for a Java session

If you enabled the detailed logging and still cannot find a solution from the messages, contact IBM Software Support for additional assistance for troubleshooting with the collected detail information.

16.10.1 Changing the log level of the system and collection logs

You can change the log level for the system log or the collection log from the administration console.

To change the log level of the system log, follow these steps:

1. Log in to the administration console, and select the **System** link on the top.
2. Select the **Log** tab with edit mode, and click the **Configure log file options** link.
3. In the Type of information to log box, change the value. Select **Error messages only**, **Error and warning messages**, or **All messages**.

   > **Maximum file size:** You can also change the maximum file size of the log file, maximum number log files, and the default locale in the log file in the same window.

4. Enter `esadmin system stopall` to stop the system, and enter `esadmin system startall` to make the change take effect.

> **Log-level changes and their impact on your system:** The more the log level increases, the more the log file size rapidly grows and rotates. When you increase the log level, consider changing the maximum file size or maximum number of the log files. Make sure that your system has enough disk space to store these log files.

When the log configuration file option is updated, the `ES_NODE_ROOT/master_config/system_log.prp` file is updated. The file does not exist until you update the configuration in the administration console.

To change the log level of the collection log, follow these steps:

1. Log in to the administration console. Select the **Collection** link on the top, and select the collection in question with the edit mode.
2. Select the **Log** tab with the edit mode, and click the **Configure log file options** link.
3. In the Type of information to log box, change the value. You can select **Error messages only**, **Error and warning messages**, or **All messages**.
4. Enter the **`esadmin system stopall`** command to stop the system, and enter the **`esadmin system startall`** command to make the change take effect.

Use this same task for the change in log levels and the impact on your system. Make sure that the system has enough disk space to store these log files.

When the configuration is updated, the `ES_NODE_ROOT/master_config/`*`collectionID`*`_log.prp` file is updated.

16.10.2 Generating a javacore or a heapdump for a Java session

Most Content Analytics sessions are run as Java processes. Sometimes, you might want to see how much the heap is used in the Java process. You can see the heap usage by generating a javacore and a heapdump for the Java session. This procedure is especially helpful when you troubleshoot out-of-memory related issues, session hang, or slow performance-related issues.

Notes about javacore and heapdump:

- How you analyze the generated javacore or heapdump is outside the scope of this book.
- You cannot issue the **`esadmin getMemStatus -dumpJava`** (or **`-dumpHeap`**) command against a session that is no longer working.
- You cannot create a javacore or heapdump for the searchapp, admin, or searchserver sessions because these sessions run under the root privilege that uses a well-known port. If you need to create the javacore or heapdump for a session, follow these steps:
  a. Comment out the `impersonate=true` line in the `ES_INSTALL_ROOT/configurations/interfaces/`*`sessionName`*`_interface.ini` file. For *`sessionName`*`_interface.ini` file, you might use `admin__interface.ini`, `searchapp__interface.ini`, or `searchserver__interface.ini`.
  b. Restart the session by using the **`esadmin`** ***`sessionID`*** **`stop`** command and then the **`esadmin`** ***`sessionID`*** **`start`** command.

To generate a javacore or a heapdump, issue either of the following commands to the sessionID in which you are interested:

```
esadmin sessionID getMemStatus -dumpJava
esadmin sessionID getMemStatus -dumpHeap
```

Depending on the command you use, the `javacore*.txt` file or `heapdump*.phd` file is generated in the `ES_NODE_ROOT/logs` directory. The file name contains the time stamp when the file is created. You can easily recognize when the file is created.

Example 16-6 shows the output when issuing the command to the Windows file system crawler session, `col_sample.WIN_77148`.

Example 16-6 The javacore output with the esadmin command

```
>esadmin col_sample.WIN_77148 getMemStatus -dumpJava
FFQC5303I IVPFile System  (sid: col_sample.WIN_77148) CCL session
exists. PID: 5288
FFQC5314I The following result occurred: <MemStatus
MaxHeapSize="536870912" TotalMemory="19181568"
FreeMemory="199224"></MemStatus>

>dir javacore*
 Directory of C:\IBM\es\esadmin\logs
```

```
04/19/2010  02:36 PM           337,419
javacore.20100419.143612.5288.0001.txt
```

As you can see, the `javacore.20100419.143612.5288.0001.txt` file is generated in the `ES_NODE_ROOT/logs` directory, which is an example for Windows. You can read the javacore file and confirm its content. Analyzing each entry in the javacore file is outside the scope of this book.

Example 16-7 shows the output when issuing the command to the Windows file system crawler session `col_sample.WIN_77148`.

Example 16-7   Heapdump output with the esadmin command

```
>esadmin col_sample.WIN_77148 getMemStatus -dumpHeap
FFQC5303I IVPFile System  (sid: col_sample.WIN_77148) CCL session
exists. PID: 5288
FFQC5314I The following result occurred: <MemStatus
MaxHeapSize="536870912" TotalMemory="19181568"
FreeMemory="10974120"></MemStatus>

>dir heapdump*
 Directory of C:\IBM\es\esadmin\logs

04/19/2010  02:40 PM         1,266,582
heapdump.20100419.144029.5288.0002.phd
```

The output file is also generated in the `ES_NODE_ROOT/logs` directory, which is an example for the Windows platform.

The heapdump file is not human readable. Analyzing the heapdump is outside the scope of this book. To see the content with visual representation, see “HeapAnalyzer: A graphical tool for discovering possible Java heap leaks” at the following web address in IBM alphaWorks:

http://www.alphaworks.ibm.com/tech/heapanalyzer

A

Security in IBM Content Analytics

This appendix provide further information about security in IBM Content Analytics. It provides information about typical situations that are related to security, especially when you use the text miner application with security. It includes a scenario of when Content Analytics uses the default web application server, Jetty. When you use the WebSphere Application Server as the web application server, go to the IBM Content Analytics Information Center at the following address and search on *enabling security in WebSphere Application Server*:

`http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

This appendix includes the following sections:

- The security concept in Content Analytics
- Enabling login security in the embedded application server (Jetty)
- Configuring application user roles
- Limiting user access to the text analytics collection

The security concept in Content Analytics

A text analytics collection contains the content extracted from the enterprise. As such, it must provide stringent safeguards to protect content from unauthorized access. Content Analytics addresses this need in several ways, each of which are explained in this section.

Security is a broad term that includes many different concepts. When reading this section, you must understand the distinction between these concepts and the relationships to each other. The most relevant concepts include authentication and authorization.

Authentication

Authentication is the process by which the system verifies that the identity of a user is who they say they are. Because access control is normally based on the identity of the user who requests access to a resource, authentication is essential to provide effective security.

Enterprises generally perform user authentication during the login process. The user enters a user ID and an associated password. The authentication process is a part of the IT infrastructure. It is typically performed by the web server or application server with a user registry (for example, Lightweight Directory Access Protocol (LDAP) repository).

Content Analytics has been engineered to work with the existing authentication components. It does not require a separate login process for the users. When Content Analytics requires the identity of the user who is logged in, it interacts with the host environment (in this case, the Jetty web server or WebSphere Application Server) or the application to obtain the user's credentials. With this approach, Content Analytics can be smoothly integrated with the existing authentication policies of an enterprise without requiring a separately maintained user registry.

Authorization

Authorization is the mechanism by which a system grants or revokes the right to access specific data or perform certain actions. Normally, a user must first log in to a system, using an authentication system as described previously. Next, the authorization mechanism controls the operations that the user might perform by comparing the user ID to an access control list (ACL). Content Analytics employs several levels of access control that can be used independently or together to provide increasing levels of authorization.

Content Analytics offers the following levels of access control:

Administrative Controls who can set up and maintain collections.

Collection Controls who has access to text analytics collections.

Document Controls who has access to which documents. This level is only supported within a search collection and not within a text analytics collection.

Encryption Encrypts sensitive data, such as passwords, specified in the administration console.

Administrative access control

Within an enterprise, multiple text analytics collections are likely to be created for different applications. Collections can contain documents from multiple data sources. Each application and its associated collection can require the expertise of different individuals in the organization to aid in the setup and configuration of the collection.

Content Analytics supports this kind of differentiation by allowing the administrator to assign individual administrator user IDs to one or more specific collections. The administrator can indicate that a particular user ID with a given role can access all collections or selected collections. Roles are an abstract logical grouping of users that are predefined by the Content Analytics product.

For administration, the following roles are predefined:

Enterprise Administrator
Has super user access to all administrative functions of the Content Analytics system. The administrator user ID provided at installation time is automatically assigned to this role.

Collection Administrator
Can edit, control, and monitor only the crawl space and properties of the collections to which they have access.

Operator Can monitor and control the collections to which they have access. Cannot edit or change properties or settings.

Monitor Can monitor the system or the collections to which they have access. Cannot control operations or edit properties.

Collection-level access control

Access to text analytics collections can be restricted to only those applications that are granted access by the administrator. The applications are those applications that are supplied by Content Analytics, such as the text miner application. They can also be the client programs that are developed by the customer using the IBM Search and Index API (SIAPI). Each application is assigned an application ID. The administrator associates one or more text analytics collections with one or more of the created application IDs.

See "Limiting user access to the text analytics collection" on page 647 for an example of how to achieve the collection level access control along with the user roles.

Document-level access control

Document-level access control ensures that users who analyze documents can only access those documents that they are authorized to see. The following primary forms of document-level access control are possible:

- Security-token assignment
- Native source security

Text analytics collection versus search collection: Document-level security is not supported in the text analytics collection. Document-level security is supported only in the search collection that does not use facets. Although it is not supported in the text analytics collection, this information is included for your reference if you need to work with the search collection that does not contain facets.

Security-token assignment

Security-token assignment is accomplished by allowing the administrator to associate one or more security tokens with each document at crawl time. The security token, except for the default token, can represent a valid operating system group ID, user ID, or any other value as determined by the administrator. By default, each document is assigned a public token so that the document can be accessed by everyone.

The following methods are available to replace the public security tokens with a different value:

- The value can be specified by the administrator through the administration console.
- The value can be extracted from an administrator-designated field in the crawled document.
- The value can be determined by a user-defined Java routine, by using the Content Analytics security token plug-in API.

  The security token plug-in API provides an entry-point to facilitate deployment and integration of the Content Analytics into an existing security infrastructure. The goal is to permit the application of business and security rules to achieve document-level security.

  The security plug-in is supported for all the crawlers except the web crawler (HTTP/HTTPS) and the NNTP. With this API, customers can write a Java routine that implements logic to gather ACLs when crawling documents. For each document fetched from the data source, the crawlers call the security token plug-in, which returns the security tokens to be associated with the document in the index.

At search and analysis time, the client application must supply one or more security tokens. If no security token is supplied, the default public token is automatically applied during the search. The security tokens stored with the documents are compared to the security tokens sent from the client application. Only those documents that match the security token specification are returned. The client application has the flexibility to specify inclusion and exclusion in the list of security tokens.

Native source security

With native source security, users have access to only those documents that a user is permitted to access as defined by the native ACLs of the source repository. This method combines the storage of native ACLs in the index with the real-time consultation of the originating repositories to determine what documents a user is allowed to see.

The storage of high-level native ACLs in the index is necessary to ensure adequate search performance but alone does not assure comprehensive document level security. The host software of the originating repository of the document becomes the final arbiter about whether the user is allowed access and thus guarantees enforcement of the native ACL of the document.

Content Analytics requests the host repository to check access to a document through a technique known as *impersonation*. Impersonation involves establishing a session with the back-end repository by using the security credentials of the user. The back-end repository works as though it interacts directly with the user and consequently responds with data that the user is authorized to access.

The technique of impersonating a user requires that Content Analytics maintains the credentials of the user to each specific back-end repository. Similar to how the user might be asked to identify themselves to the original enterprise repositories to perform an action, Content Analytics requires the same credentials for any documents in the index that require authorization. The multiple security credentials of a user are prompted for and stored by Content Analytics in a user registry. The information is then retrieved when the user establishes a session in the text miner application.

By not knowing in advance the types of results that might be generated from text mining (and then the back-end repositories that will be impersonated), all security credentials must be supplied on each individual search. In addition to the security credentials, the groups to which the user belongs must be provided.

Data encryption

A fourth area of security employed by Content Analytics is data encryption. During the administration of a text analytics collection, instances exist where the administrator is asked to provide sensitive information such as passwords used by the crawlers to access the back-end data sources. All supplied passwords are masked when they are entered or displayed in the administration console.

In addition, the passwords are further encrypted by using industry-proven techniques before being stored in the configuration database. Encryption provides additional protection of the passwords even if access is gained to the Content Analytics machines.

The content stored in a text analytics collection is not encrypted. The reason is primarily due to performance concerns because it might be too time consuming to decrypt content during a typical analytic session. A text analytics collection is in a form that supports high-speed text analysis and discovery and is not easily discernible by a person.

Enabling login security in the embedded application server (Jetty)

When you install the Content Analytics with the Jetty web application server, the security options, such as user authentication and the collection-level security mechanism, are not enabled by default. If you want to use the security features in Content Analytics, you must enable security in the Jetty web application server. This section explains how to enable security in the Jetty web application server.

WebSphere Application Server: If you use WebSphere Application Server instead of the Jetty web application server to deploy the text miner application, follow the steps to enable security in WebSphere Application Server in the product manual.

For WebSphere Application Server 7, see the WebSphere Application Server Version 7.0 Information Center at the following address:

`http://publib.boulder.ibm.com/infocenter/wasinfo/v7r0/index.jsp`

After you enable the security in WebSphere Application Server, follow the following steps:

1. Issue **`eschangewaspw`** command. For the detailed steps, go to the IBM Content Analytics Information Center at the following address and search on *Enabling security in WebSphere Application Server*:

   `http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp`

2. Enable security in the Jetty web application server as explained in this section. Use the same LDAP user registry that you used in WebSphere Application Server security configuration.

We do not address this topic further, because it is outside the scope of this book.

To enable security using the Jetty web application server, follow these steps:

1. Log in to the administration console as the administrative user, and select the **Security** tab (Figure A-1).

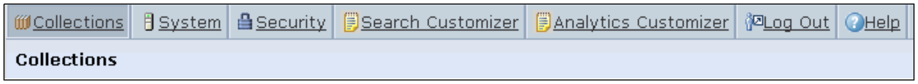


Figure A-1 Selecting the Security tab in the administration console

2. Select the **System Security** tab (Figure A-2), and select **Configure embedded application server login settings**.

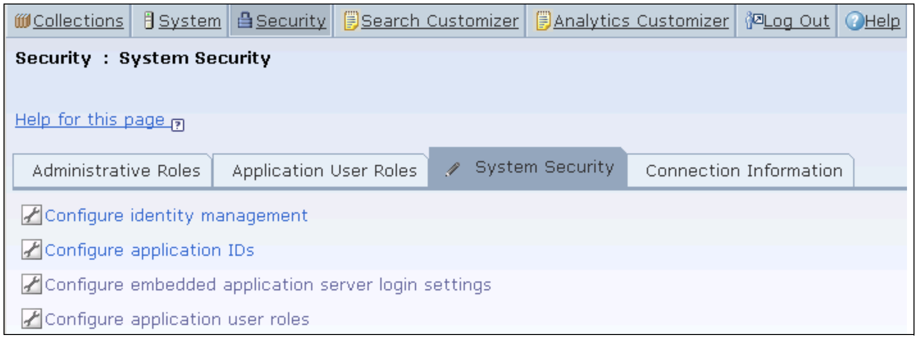
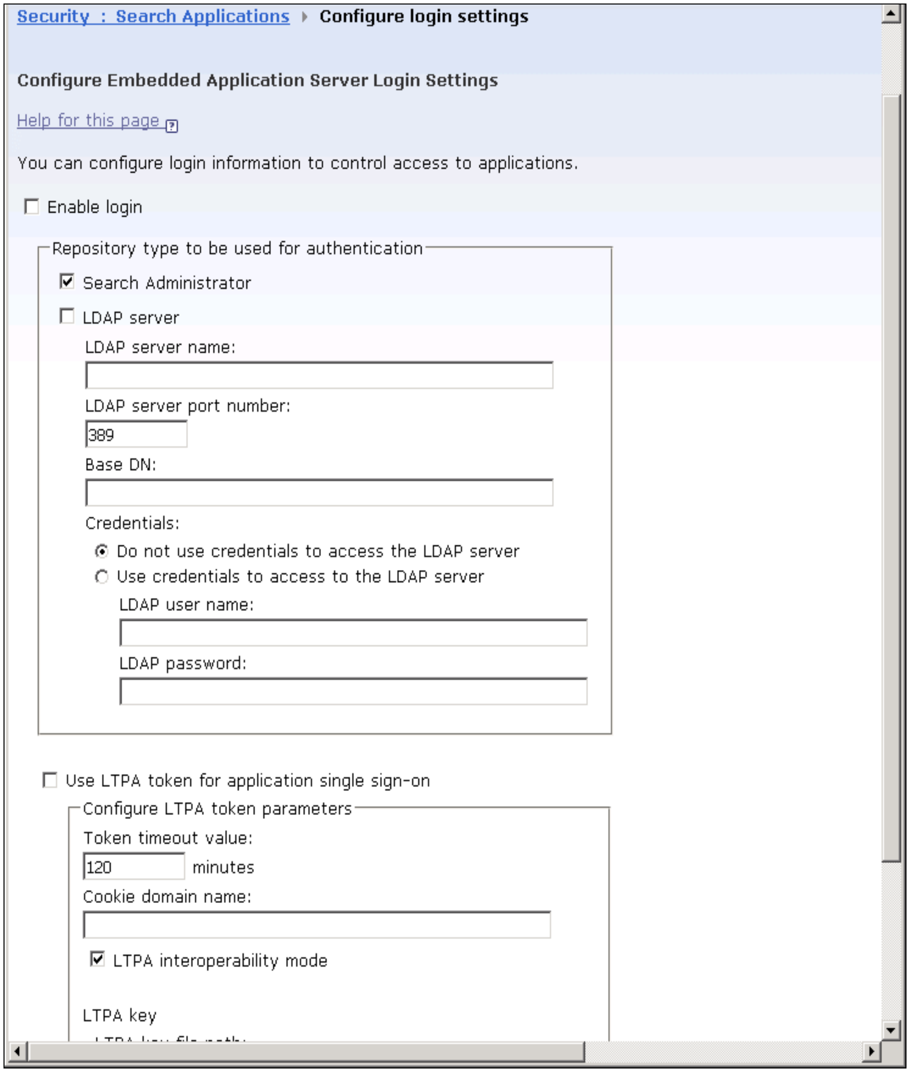


Figure A-2 Configuring the embedded application server login settings

3. In the Configure Embedded Application Server Login Settings panel (Figure A-3), select the **Enable login** check box, and select the **LDAP server** check box to enable the user authentication. Provide the necessary information such as the LDAP server name and Base DN. When you use LTPA token for application single sign-on, specify the LTPA key file location.



Security : Search Applications ▸ Configure login settings

Configure Embedded Application Server Login Settings

Help for this page

You can configure login information to control access to applications.

□ Enable login

Repository type to be used for authentication

☑ Search Administrator

□ LDAP server

LDAP server name:

LDAP server port number: 389

Base DN:

Credentials:

● Do not use credentials to access the LDAP server

○ Use credentials to access to the LDAP server

LDAP user name:

LDAP password:

□ Use LTPA token for application single sign-on

Configure LTPA token parameters

Token timeout value: 120 minutes

Cookie domain name:

☑ LTPA interoperability mode

LTPA key

LTPA key file path:

Figure A-3 Configuring the Jetty login setting on the administration console

Hints and tips:

- How you set Base DN depends on the LDAP server you use. Consult with your LDAP server administrator accordingly.
- Select the **Search Administrator** check box when you enable the login settings for the first time. By selecting this check box, you can always log in as the administrative user that you set during the installation. This setting helps you to change the configuration if any problems occur during the LDAP configuration.

4. After you save the change, log off from the administration console, and issue the following command to restart the system:

```
esadmin system restart
```

Message on the administration console: You can restart the system by issuing the **`esadmin system stop`** and **`esadmin system start`** commands. However, you can use the **`esadmin system restart`** command instead of issuing both of the **`start`** and **`stop`** commands.

Configuring application user roles

When you use the text miner application to analyze your data, you can grant various user roles access to functions in the text miner application. When configuring application user roles, you can configure the application user roles for the system (used by default), or configure the application user roles for specific users or groups. This section explains how you configure application user roles.

Configuring application user roles for the system

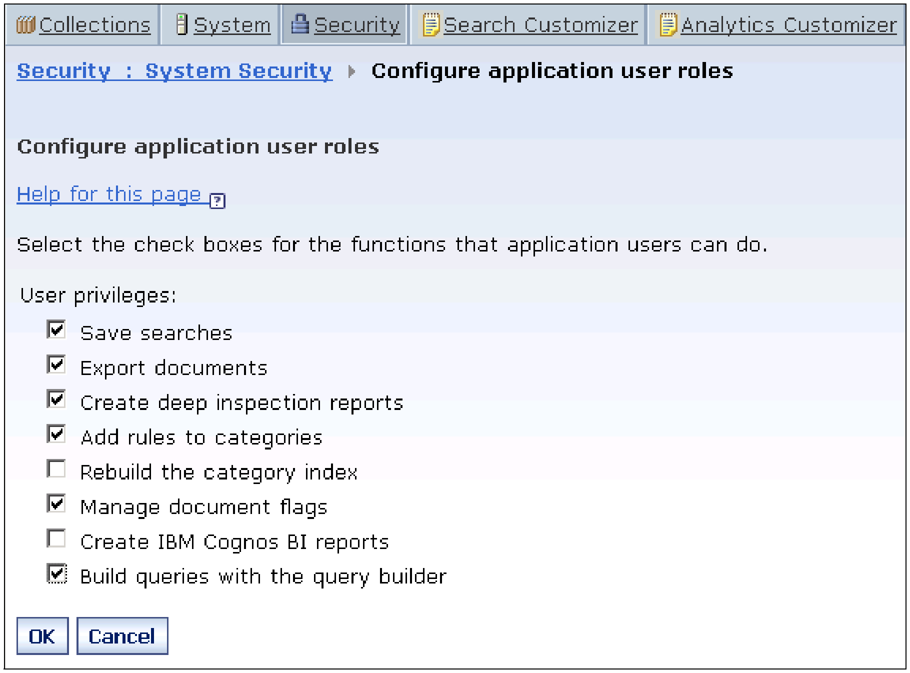
When the application user roles for the system are configured, they are used in the following situations:

- The security is not enabled in the Jetty web application server.
- The application user roles are not configured for specific users or groups and security is enabled in the Jetty web application server.

To configure application user roles for the system, follow these steps:

1. Log in to the administration console as the administrative user.
2. Select the **Security** tab.

3. Select the **System Security** tab, and select **Configure application user roles**.
4. In the Configure application user roles panel (Figure A-4), select the user privilege check boxes that corresponds to the functions that application users can perform. Click **OK** to save the changes.



Configure application user roles

Configure application user roles

Select the check boxes for the functions that application users can do.

User privileges:

☑ Save searches
☑ Export documents
☑ Create deep inspection reports
☑ Add rules to categories
□ Rebuild the category index
☑ Manage document flags
□ Create IBM Cognos BI reports
☑ Build queries with the query builder

Figure A-4 Configuring application user roles for the system

5. Log off from the administration console.
6. Issue the following command to restart the system:

```
esadmin system restart
```

The following user privileges are some examples of the privileges that can be enabled:

- Query builder. For further details about how the query builder works, see 5.4, “Query builder” on page 180.
- Cognos BI integration. For further details about creating a Cognos BI report, see Chapter 13, “Integrating Cognos Business Intelligence” on page 525.
- Document flagging. For further details about working with document flags, see 5.7, “Document flagging” on page 205.

Configuring the application user roles: The icons associated with the feature are available in the text miner application when the feature is enabled.

Configuring application user roles for specific user or group

When you enable security in the Jetty web application server as mentioned in the previous section, you can configure the application user roles for a specific user or group.

To configure the application user roles for specific users or groups, follow these steps:

1. Log in to the administration console as the administrative user, and click the **Security** tab.
2. Click the **Application user roles** tab (Figure A-5), and click **Add User or Group**.

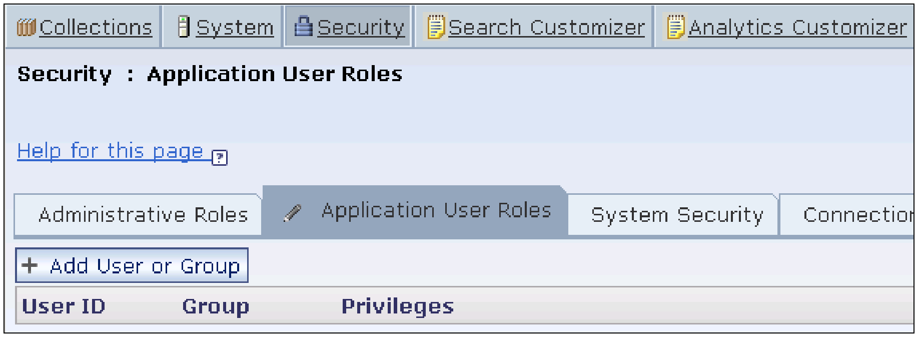
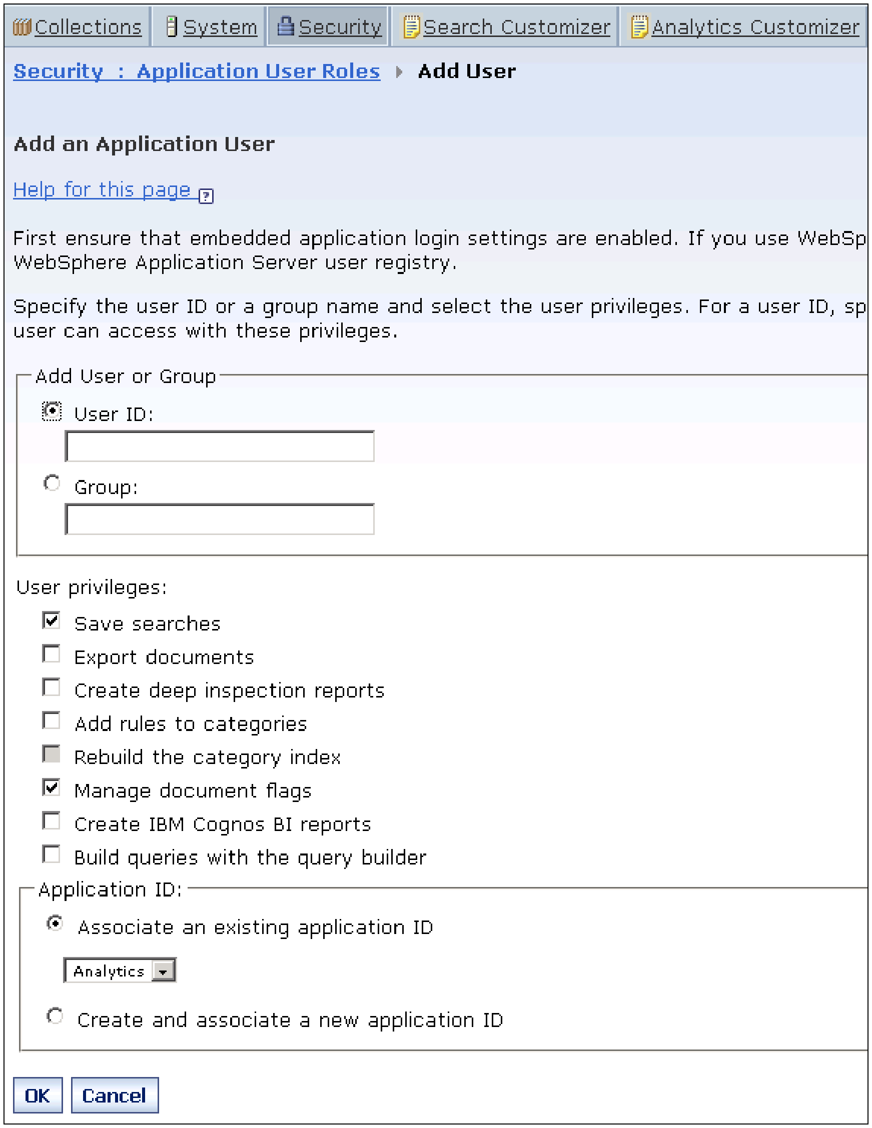


Figure A-5 Configuring the application user roles for a specific user or group

3. In the Add an Application User panel (Figure A-5), under Add User or Group, select the **User ID** or the **Group** radio button. Type the user name in the User ID field or the group name in the Group field. Select the necessary application user role check boxes that you want assigned to the specified user or group. The click **OK**.

– When you select the **User ID** radio button, you see the options in the panel as shown in Figure A-6.



Security : Application User Roles ▸ Add User

Add an Application User

Help for this page

First ensure that embedded application login settings are enabled. If you use WebSp
WebSphere Application Server user registry.

Specify the user ID or a group name and select the user privileges. For a user ID, sp
user can access with these privileges.

Add User or Group

● User ID:

○ Group:

User privileges:

☑ Save searches
☐ Export documents
☐ Create deep inspection reports
☐ Add rules to categories
☐ Rebuild the category index
☑ Manage document flags
☐ Create IBM Cognos BI reports
☐ Build queries with the query builder

Application ID:

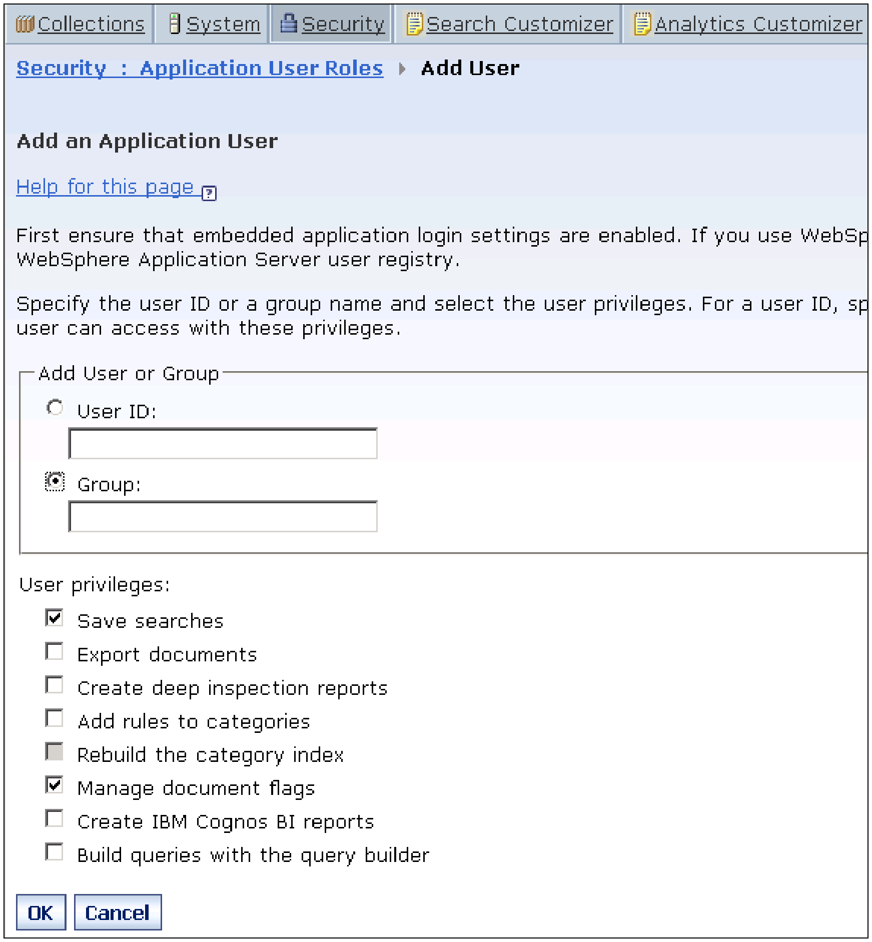
● Associate an existing application ID

Analytics

○ Create and associate a new application ID

Figure A-6 Configuring the application user roles for a user ID

– When you select the **Group** radio button, you see the options in the panel as shown in Figure A-7.



Security : Application User Roles ▸ Add User

Add an Application User

Help for this page

First ensure that embedded application login settings are enabled. If you use WebSp
WebSphere Application Server user registry.

Specify the user ID or a group name and select the user privileges. For a user ID, sp
user can access with these privileges.

Add User or Group

○ User ID:

● Group:

User privileges:

☑ Save searches
□ Export documents
□ Create deep inspection reports
□ Add rules to categories
□ Rebuild the category index
☑ Manage document flags
□ Create IBM Cognos BI reports
□ Build queries with the query builder

Figure A-7 Configuring the application user roles for a group

4. After you saved the change, log off from the administration console.
5. Issue the following command in a command prompt as the Content Analytics administrator to restart the system:

   ```
   esadmin system restart
   ```

Assigning user roles to the user or the group is only used when you explicitly assign the user role for a specific user or group. As explained in the previous section, if the user or group is not explicitly granted for a particular user role, the default system configuration is used.

Limiting user access to the text analytics collection

When you use the text miner application to analyze your data, you can limit the specific text analytics collection that a user can access to allow them to view and analyze it. This section explains how you can do task by using the Jetty web application server.

Jetty as a web application server: The following steps apply to Content Analytics using Jetty as a web application server. When you use WebSphere Application Server as the web application server, you must use the security mechanism for WebSphere Application Server. For more information, see the WebSphere Application Server Version 7.0 Information Center at the following address:

`http://publib.boulder.ibm.com/infocenter/wasinfo/v7r0/index.jsp`

Task overview

Allowing role-based access to the text analytics collection within the text miner application entails the following tasks:

1. Mapping an application ID with collections and the user
2. Verifying security privileges

You perform these operations from the administration console. After the security configuration is modified, restart Content Analytics.

Scenario overview

The following scenario is used to demonstrate the steps required to implement security control:

- You are the administrative user of a Content Analytics system and need to set up role-based access control with the text miner application.
- You must set up the Content Analytics to meet following requirements:
  - *user1* is the sales manager and needs to access the collections that are associated with the application ID called `Sales Report Collections`.
  - *user2* is the customer support manager and needs to access the collections that are associated with the application ID called `Customer call Report collections`.

– Users user1 and user2 are different users and must not be able to see the collection of the other person. That is, user1 cannot see the data of user2 and vice versa.

Mapping an application ID with collections and the user

To assign the application user roles, follow the steps in “Configuring application user roles for specific user or group” on page 644. Then continue with the following steps:

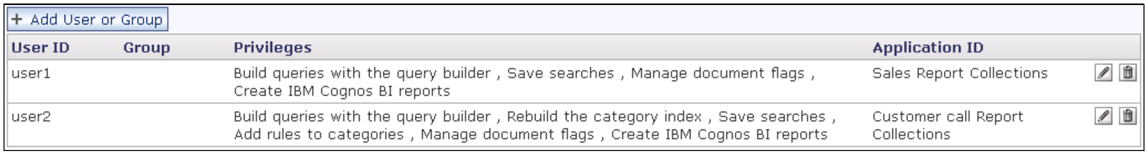
1. Log in to the administration console as the administrative user.
2. Select the **Security** tab.
3. Select the **Application user roles** tab, and click **Add User or Group**.
4. In the Add an Application User panel, complete these steps:
   a. Under Add User or Group, select **User ID** or **Group** radio button, and enter the user name or group name.

      For this scenario, select the **User ID** radio button to assign the application user role and application ID.

   b. Select the **Create and associate a new application ID** radio button in the Application ID section, and select the collection name check box to associate with the application id.

      Configuring an application ID: If you have configured the application ID already, you can select **Associate an existing ID** radio button instead.

   c. After you complete all the information, click **OK**.
5. Add other users with other roles, if necessary. In this example, add the application role and application ID mapping for user2 (Figure A-8).



User ID	Group	Privileges	Application ID
user1		Build queries with the query builder , Save searches , Manage document flags , Create IBM Cognos BI reports	Sales Report Collections
user2		Build queries with the query builder , Rebuild the category index , Save searches , Add rules to categories , Manage document flags , Create IBM Cognos BI reports	Customer call Report Collections

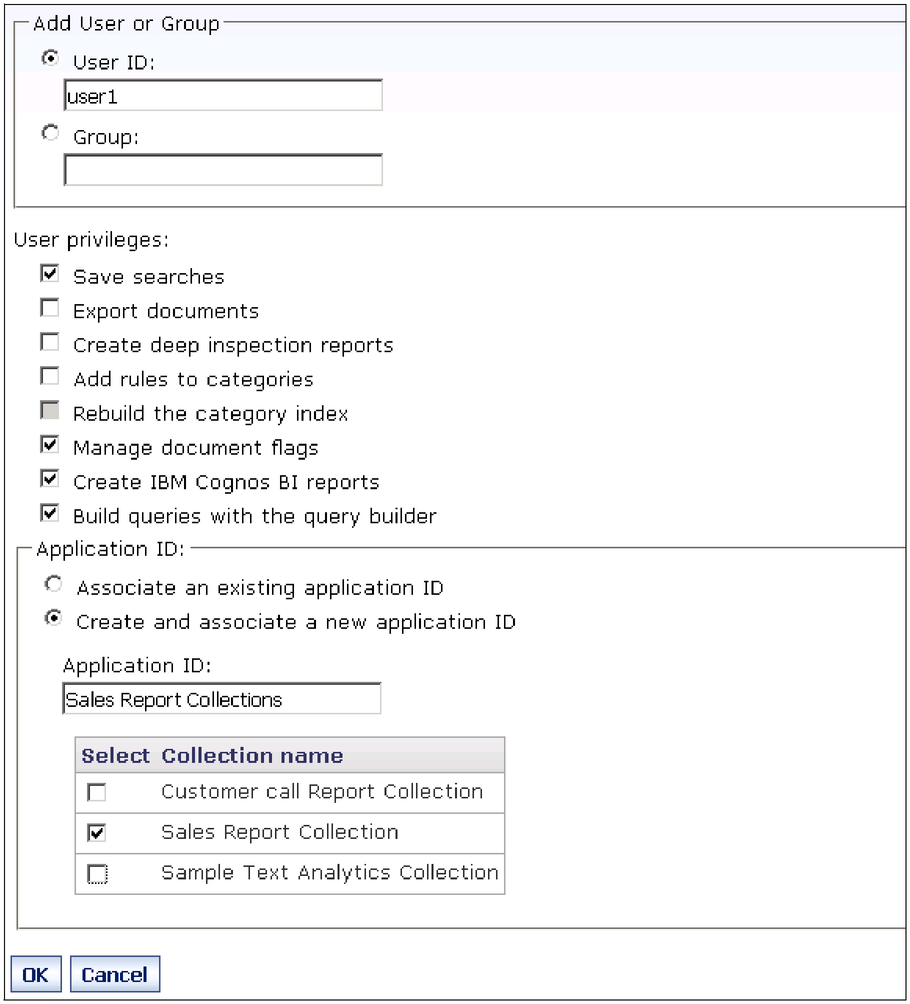
Figure A-8 After the user is added with application user role and application ID

6. After you save the change, log off from the administration console, and issue the following command to restart the system:

```
esadmin system restart
```

In this scenario, we associate the application ID and the collections as follows:

- Assign the `Sales Report Collections` application ID to the `Sales Report Collection` collection, and associate it with user1 as shown in Figure A-9.



Add User or Group

● User ID: user1

○ Group:

User privileges:

☑ Save searches
☐ Export documents
☐ Create deep inspection reports
☐ Add rules to categories
☐ Rebuild the category index
☑ Manage document flags
☑ Create IBM Cognos BI reports
☑ Build queries with the query builder

Application ID:

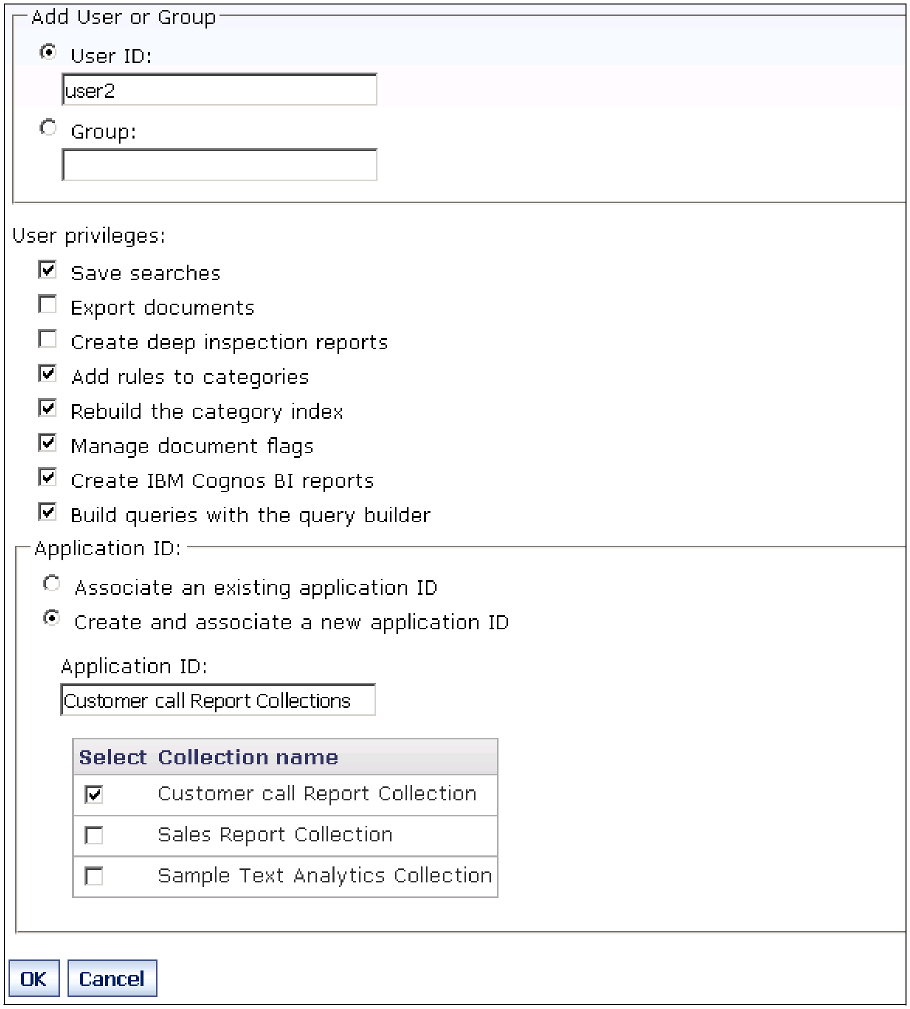
○ Associate an existing application ID
● Create and associate a new application ID

Application ID: Sales Report Collections

Select	Collection name
☐	Customer call Report Collection
☑	Sales Report Collection
☐	Sample Text Analytics Collection

Figure A-9 Application ID and role Mapping information for User1

- Assign the `Customer call Report Collections` application ID to the `Customer call Report Collection` collection, and associate it with user2 as shown in Figure A-10.



Add User or Group

● User ID:

user2

○ Group:

User privileges:

☑ Save searches

□ Export documents

□ Create deep inspection reports

☑ Add rules to categories

☑ Rebuild the category index

☑ Manage document flags

☑ Create IBM Cognos BI reports

☑ Build queries with the query builder

Application ID:

○ Associate an existing application ID

● Create and associate a new application ID

Application ID:

Customer call Report Collections

Select	Collection name
☑	Customer call Report Collection
□	Sales Report Collection
□	Sample Text Analytics Collection

Figure A-10 Application ID and role mapping information for User2

Verifying security privileges

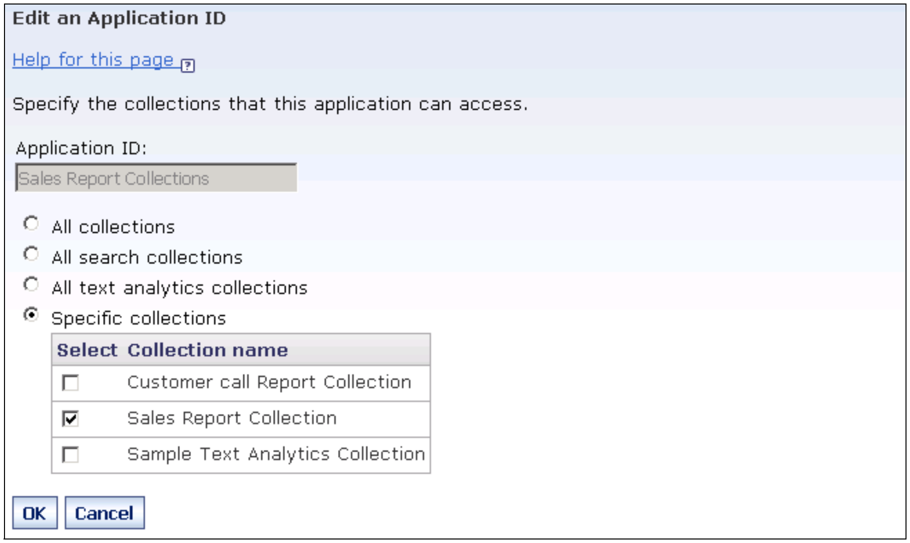
After you restart the system, try to log in as the user that is assigned a role, and verify that the user only views the collections that are associated with their roles. You view and select the collections that the user has authority to analyze from **Preferences** in the text miner application.

In this scenario, when you log in with the user1 user ID, only the *Sales Report Collection* collection is listed and selectable in the Preferences panel. If you log in with the user2 user ID, only the *Customer call Report Collection* collection is listed and selectable in the Preferences panel. If you can verify the result as explained, you have set up role-based access security successfully in the Jetty web application server.

Managing an application ID

After you define the mapping between an application ID and collections, you can manage the application ID from the administration console by using the following steps:

1. Log in to the administration console as the administrative user.
2. Select the **Security** tab.
3. Select the **System Security** tab, and click **Configure application IDs**.
4. Either click **Add Application ID** to create an application ID, or click the **Edit** icon associated with an existing application ID to edit it. In this scenario, we click the **Edit** icon associated with the Sales Report Collection collection.
5. In the Edit an Application ID panel (Figure A-11), modify the application ID based on your needs. Then click **OK** to save the changes.



Edit an Application ID

Help for this page

Specify the collections that this application can access.

Application ID: Sales Report Collections

○ All collections
○ All search collections
○ All text analytics collections
● Specific collections

Select	Collection name
□	Customer call Report Collection
☑	Sales Report Collection
□	Sample Text Analytics Collection

Figure A-11 Editing an existing application ID and collection mapping

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The IBM Redbooks publication *IBM Classification Module: Make It Work for You*, SG24-7707, provides additional information about the topic in this document.

You can search for, view, download or order this document and other Redbooks, Redpapers, Web Docs, draft, and additional materials, at the following website:

ibm.com/redbooks

Other publications

The following publication is also relevant as a further information source:

- *IBM Cognos Content Analytics, Version 2.1.0 Administration Guide*, SC19-2875

Online resources

These websites are also relevant as further information sources:

- IBM Content Analytics Version 2.2 Information Center

  http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp
- IBM Content Analytics latest supported data sources

  http://www.ibm.com/support/docview.wss?rs=4173&uid=swg27015094
- IBM Content Analytics latest system requirements

  http://www.ibm.com/support/docview.wss?rs=4173&uid=swg27015092
- IBM Content Analytics support

  http://www.ibm.com/support/docview.wss?rs=4173&uid=swg27015096

- IBM Classification Module information

  http://www.ibm.com/software/data/content-management/classification

- Collecting data for InfoSphere Classification Module

  http://www.ibm.com/support/docview.wss?uid=swg21417244

- Content Assessment information

  http://www.ibm.com/software/data/content-management/assessment.html

- ECM search and discovery product information

  http://www.ibm.com/software/data/content-management/products/search.html

- IBM Archive and eDiscovery Solution Information Center (for Content Collector, eDiscovery, and eDiscovery Manager)

  http://publib.boulder.ibm.com/infocenter/email/v2r1m1/index.jsp

- IBM Classification Module information center

  http://publib.boulder.ibm.com/infocenter/classify/v8r7/index.jsp

- IBM LanguageWare information

  http://www.alphaworks.ibm.com/tech/lrw/

- Technote “Building a knowledge base for IBM Classification Module V8.7”

  http://www.ibm.com/support/docview.wss?uid=swg27015916

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

A

access control list (ACL) 17
access control, document-level 636
accuracy
  classification 363
  textual data 51
ACL (access control list) 17
actionable insight 3, 46, 143, 280, 395
adic.xml file 308
administration console 16
  access 85
  log 608
administrative user 76
advanced option 574, 577
advanced search 162
advanced techniques 480
advanced text analytics 453
  function 450
aggregate rules 457
AIX system monitoring 602
analysis
  activities 325
  car complaint data 337
  complaint data 322
  cycle 53
  data, multiple viewpoints 322
  insight from various viewpoints 323
  limited scope 153
  of content 2
  of customer contact record 66
  of large collections 593
  typical cycle 53
  viewpoints 326
  voice of customer 60
  weather 323
  with text miner application 298
analytics of textual data 2
analyzable field 32
analyzed document 398
AND NOT operator 168
AND operator 166
annotation 457
annotator 18, 32, 450
  custom 453, 468, 475
    development 468
  Dictionary Lookup 19, 35, 299, 302, 319, 452
  export 458
  IBM Classification Module 384, 452
  Named Entity Recognition 452
  Pattern Matcher 19, 35, 309–310, 452
  Regular Expression 469
  UIMA compliant 458
  UIMA pipeline 19
  validation techniques 482
Apache Lucene indexer 20
Apache UIMA SDK 468
application ID 636, 647
application toolbar 147
application user role 180, 194, 527
architecture 15
assessment process 522
association
  decision plan with collection 380
  keyword to a facet 303
  pattern matching with a facet 292
authentication 634
authorization 634
auto manufacture 29
automation of classification 358
average document size 573

B

backend repository 637
backup
  collection configuration file 140
  entire index 140
bar chart 222, 228
  show all charts 233
  show selected charts 233
battery failure 432
behavior of IBM Content Analytics 476
binary content 397, 399, 403, 507
bird's eye view 247
bottleneck 581
build collection 37
building, Optional Facet Index 94

business data preservation and exploitation 507
business requirement 490, 558
business value 484

C
call center 28, 54, 65, 280, 287, 291
  improvement 66
  record 28
CAS (Common Analysis Structure) 19, 399, 404, 468
  debugging 481
  view 468
CAS Visual Debugger (CVD) 481
catalog.xml file 363, 367, 402, 426
categorization type, document clustering 343
categorized documents, working with 353
category 5, 90, 395
  based on document clustering 351
  rebuilding 201
  rule 198
  score 381
  searching within 359
  tree, access 195
CCL (Common Communication Layer) 5
character rule-based annotation 456
character rules 456
Charts per page indicator 233
CK property value 613
classification 46, 358
  accuracy 363
  automate 358
  automation 358
  distribution 48
  document 452
  technology 359
Classification Module 30, 35
  import of XML files 355
  technology 359
  training a knowledge base 378
Classification Workbench 383
  report 383
cluster 353
  proposed 348
  refining 348
cluster proposal, creation 346
clustering
  document clustering 351
  preferred practices 356
Cognos 8 BI 525
  database connectivity 528
Cognos BI server URI 531
Cognos SDK 526
collaboration crawler 17
collection 5, 11, 21, 27, 32–33, 36–37, 572
  analyzing large collections 593
  associating a decision plan 380
  configuration file backup 140
  creation 451
  dynamic collection 406
  exported content 508
  ID 574, 599
  information 628
  log 608–609, 629
  memory consumption 572
  number of documents 572
  text analytics 31, 90, 638
Collections view 127
column 32
  name 152
commit interval 587
Common Analysis Structure (CAS) 19, 399, 404, 468
  debugging 481
  view 468
Common Communication Layer (CCL) 5
complaint data analysis 337
concept-based searching 35
conceptual resemblance 364
conceptual search 364, 381
config_ID 575
configuration 73
  custom export plug-in 624
  decision plan 378
  deployment 139
  export 344, 409, 421
  export option in the collection 625
configuration file backup 140
Connections view 250
  useful features 252
  using 261
consumption, memory 572
content
  categorization 492
  crawling, parsing, and indexing 493
  exported, collection 508
  extraction 576
  ingesting 492

reindexing 382
content analysis 2, 9, 21, 45–46, 279, 357
fine-tuning 361, 452
frequency and deviation 49
IBM Classification Module 357
improvement 452
performing 69, 279
scenarios 287
Content Analytics 60
advantages 45
architecture 15
data model 33
differences with Content Analyzer 5
features 34
import feature 396
installation 36
major components 16
mechanical operation 9
power 51
security features 639
successful use cases 64
text analytics 33
use-case scenarios 28
call center 28
insurance fraud 28
quality assurance 29
working environment 69
Content Assessment 5
Content Collector
ingesting content 492
resource preparation 508
Content Collector for Email 523
content decommissioning 2, 484–485, 493, 500
scenario 487
use case 500
workflow 491
conventional metadata 399
corpus 11, 49, 360
correlation 14–15, 63, 225, 285
value 14, 24, 40, 55, 222, 245, 248–249
coverage 51
crawl space 17
crawl type selection 127
crawler 16, 39, 123, 397, 572, 578, 614, 619, 621–622, 638
categories 16
collaboration 17
component 129
configuration 397
custom plug-in 132
email 17
enterprise data source 16
file system 17, 281
monitoring the status 597
relational database 17
running with parser and indexer 495
scheduling 127
session 622
source type 32
starting using the administration console 129
status check 131
stopping using the administration console 130
web crawler 637
web-based 16
WebSphere portal 17
Windows file system 123
crawling
content 493
data, methods of 397
document 620
schedule 622
create dictionary 339
CSV file import 388
CSV format 339
custom annotation 458
custom annotator 453, 468, 475
creating and usage of 469
development 468
use cases 455
character rule-based annotation 456
custom annotation 458
dictionary-based annotation 455
rule annotation 457
validation techniques 482
custom crawler plug-in 132
custom dictionary 280, 283, 286, 289, 297, 300
adding 300
adding nouns 300
candidate words 298
custom export plug-in configuration 624
custom Java export plug-in 22
custom plug-in 446, 625
for deep inspection 447
custom rule 311
customer comment 2
customer contact
center 49, 64
record 4, 49, 61, 64

record analysis 66
customized dictionary 326, 338
CVD (CAS Visual Debugger) 481
cycle of analysis 53

D

Dashboard
charts as images 277
layout 262
configuration 263
working with 275
Dashboard view 262
data
date-sensitive 34
directory 74
flow 22
inspection 508
manual analysis 46
model 33
database 5
connectivity with Cognos 8 BI 528
mapping file 408
table 32
URL 529
date facet 155
date field 10, 32, 34
date range facet 116
date-sensitive data 34
DB2 table 17
decay factor 237
decision plan 355, 360, 363, 371, 383–384, 508, 521
association with collection 380
configuration 378
creation of 371, 383
deployment 375
troubleshooting 626
tuning and maintaining 383
decommissioning content 485, 493, 500
workflow 491
decommissioning unnecessary information 486
deep inspection 23–24, 246, 249, 387, 396, 426, 431–432
analysis 396, 432, 441, 444
calculation 436
custom plug-in option 447
enabling 437
exported data 433
exported results 24, 432
feature 433
feature export 447
function 24, 432
icon 23, 432, 438–439
next scheduled run 443
options 434
output 445
report 432–434
generation 445
submitting a request 438
request 432–433, 444
options 434
validation 444
result 24, 447
run 432, 441
scheduling 441
default field 32
default query 164–165, 170, 172–173
default taxonomy cache type 585, 587
default view 468
density 248
deviation 15
index score 240, 242
Deviations view 238, 243
deviation index score 240
features 239
frequency of selected time period 240
diagnose 480
dictionary 57
building 309
candidates 482
creation 339, 342
custom 280, 283, 286, 289, 297, 300
customized 326
definition 290
migration 308
top-level element 308
Dictionary Editor 299, 301, 303, 305–307, 309
Dictionary Lookup annotator 19, 35, 299, 302, 319, 452
when to use 299
dictionary-based annotation 455
discovery
insights 361
trends 432
disk I/O 595, 602
disk space, temporary 610
distribution of high level classifications 48

doc_product field 282
document
  analysis 398
  body 31
  categorization deployment 352
  clustering 194, 343, 351, 354, 356
    categorization type 343
    deploy to a category 351
    granularity choice 356
    workflow 343
  content 18, 20, 221, 357, 364, 487, 498
  crawling 620
  flag 205
    add to a single document 211
    associate with 209
    authority to manage 208
    setting 209
    viewing 213
  identifier 32
  link 559
  number of documents per collection 572
  processing 32, 450, 582
  processor 11, 16, 18, 20–21, 24, 572–573, 579–580
    default number 580
    effective number 579
    maximum heap size 573
    server 595
    text analysis 23
    total number 582
  size 573
document processor threads 580
DOCUMENT_FACT table 419
document-level access control 636
Documents view 219, 290
  URL link 559
domain context 330
Domino database 33
dynamic analysis 485, 491, 502
dynamic collection 406, 485, 490–491, 495, 503

E
eDiscovery readiness 485
email archiving 523
email crawler 17
encryption 638
enterprise data source crawler 16
Entity facet 334
environment 607
  variable 574
ES_INSTALL_ROOT 574
ES_NODE_ROOT 574
esadmin check command 575
esadmin command 482, 596, 621
  output 597
  utility 596
esadmin export command 141
esadmin import command 141
esadmin utility 575, 599
esbackup command 140–141
esrepcog command 526
esrestore command 141
esviewauditlog command 609–610
esviewlog command 609
ETL (extract, transform, and load) 34
Event viewer 611
example
  alternative expressions 53
  AND NOT operator 168
  AND operator 166
  apply search operator 174
  capturing data 52
  correlation 14
  customer contact record 61
  facet 12
  keywords 156
  OR operator 171
  PC help center 61
  precision versus recall 51
expected value 242
exploiting business data 507
export 494, 507
  annotator 458
  configuration 344, 409, 421
  deep inspection results 24
  Java export plug-in 22
  keywords 339
  monitoring 408, 413, 426, 430
  option configuration and change 625
  point 22
  request 406, 419
    status 407
  search result 423
  validation of request 426, 430
Export icon 409
export option in the collection 625
Export utility 505

exported content collection 508
exported deep inspection data 433
exporting 482
  analyzed documents to relational database 415
  CAS to file system 404
  crawled documents to file system 403
  data 395–396, 400, 499
    considerations for 408
    rationale 395
  metadata to file system 403
  points 396
  reasons 415
  search results to the file system 421
  validate 426, 430
expressions, problems with 331
external hook 385
extract, transform, and load (ETL) 34
extracted text 399
extracting
  content 576
  metadata 360

F
facet 3, 5, 12, 32, 289, 359, 378
  association with
    keyword 303
    pattern matching 292
  cloud 562
  creation of 90
  example 12
  index, optional 591
  mapping to search field 90, 106
  path 292
  path used in query 156
  populate 449
  population of 449
  Predicate 334
  search 359
  selected 55, 145, 218, 244, 303, 434
  sudden increases 231
  tree 289, 292
Facet Navigation pane 148–149, 153
  Product facet 223
  Verb facet 291
Facet Pairs view 56, 244–245
  correlation values 248
  example showing keywords 435
  usage of 249
facet querying performance 593
facet tree changes 113
Faceted Path search 156
Facets view 153, 222, 225, 308, 342
  default sort order 223
  result changes 167
  synonyms 303
FAQ database 64
  modified entries 64
feedback 2
field 32
  analyzable 32
  default 32
  free-form 2
  mapping 41, 281, 381
  name 39, 281, 402, 405, 410, 414
  search 32
file extension preservation 405
file system 5, 17, 22, 37, 395–396, 399–400, 402, 485, 493, 508–510, 559
  crawler 17, 123, 281
  IBM Content Collector 508
  metadata 494, 510
  XML file 24
filter document 364
First Steps program 280
  starting 82
flag, association with selected documents 212
format error 59
formatted log 610
free-form field 2
frequency 13, 15, 63, 225
  count 13, 24, 49–50, 222, 236
    time period 235
    total number 242
  highest 235
  keyword 242
frequent keyword 223, 245, 435–436
FSC Associate Metadata task 512
FSC Post Processing task 515, 519
fuzzy search 156

G
global security 162
goal of the analyst 48
granted access 636
graphic user interface (GUI) 5
Grid view 246

GUI (graphic user interface) 5

H

heap size
  changing 578, 581, 586
  indexer, changing 586
  search instance, changing 593
heapdump 629, 631
Help link 147
high correlation 432
highest frequency 235
highest index 235
highly correlated 285

I

IATEMPDIR 610
IBM Classification Module 320, 357–358, 395, 402, 485–486, 498
  annotator 363, 384, 452, 497
  creation and deployment of resources 365
  deletion candidates 497
  server, starting 365
  XML export data 383
IBM Content Analytics 27, 30, 45, 144, 280, 357, 483, 571, 633
  behavior 476
  Classification Module annotator 358
  discovery capability 395
  export 507
  export features 387
  Help link 147
  installation 77
  installation process 72
  starting 85
  stopping 86
  text miner application 523
  user interface 143, 217
IBM Content Analyzer dictionary migration 308
IBM Content Assessment 30, 483–484
  analytics capabilities 486
  content decommission 493
    workflow 491
  Customer Metadata List 510
  major use cases 502
  offering 483
  tips 523
IBM Content Collector 30, 395, 397, 485, 491–492
  task route 523
IBM FileNet P8
  collected data 522
  creating the document task 513
ICM_NVP key 426–427
identification
  information 486
  technical trend 68
impersonation 637
implementation of security control 647
import
  category score 381
  CSV file function 394
  feature in Content Analytics 396
  history task 394
increase indicator 236–237
incremental export 406
index 20
  backup 140
  building 591
  content 493
  highest 235
  latest 235
  rebuilding 294, 573
  when to rebuild 137
index commit interval 587
indexer 16, 572
  changing maximum heap size 586
  component 581
  running with crawler and parser 495
  thread 577
    changing 582
    default number 582
    recommended number 582
    setting 582
indexer threads 582
indexing component 21
information
  collection 628
  identification 486
ingest, content 492
insight 3
  actionable 3, 46, 280, 395
  discovery 361
  from textual data 279
  search and discovery 152
  text miner application 143, 280
  without customization 283
installation 36
  directory 74

IBM Content Analytics 72, 77
verification 82
installation-related log file 608
insurance fraud 28
analysis 28
intent of words 358
interval estimation 249
investigation, problem 480
iostat commands 602

J
Java session 629
javacore 629
JavaScript widget, customization 564
JDBC driver name 529
Jetty web application server 647

K
keyword 10, 57, 154, 245, 382
association to a facet 303
create dictionary 339
creation of 301
distribution 339
make a better list for dictionary creation 341
synonym 302
knowledge base 355, 359–361, 395, 495, 500, 523
creation of 366, 383
deployment 375
training 378, 383
troubleshooting 626
tuning and maintenance 383

L
LanguageWare Resource Workbench 453–454
large collections analysis 593
latest index 235
LDAP server
administrator 642
check box 641
name 641
lemma 451
lexical analysis 451
limiting
scope of analysis 153
user access 647
linguistic analysis 451
Linux system monitoring 602
location of user data 574
log 608
collection log 608–609
content
understanding 609
view 609
file 605
installation-related 608
maximum size 610
formatted 610
level
changing 628
changing for collection log 629
collection log 629
system log 628
system log 608–609
changing log level 628
text miner application 608
logical document 32
logical hierarchical structure 163
lrw-view 468
Lucene document 19, 21

M
main component 571
maintenance of decision plan and knowledge base 383
maintenance record 29
manual analysis of textual data 46
mapping
file 415
for exporting analyzed documents 415
native field to search field 89, 103
search field to facet 90, 106
matching 360
max_heap property 579, 581, 586
maximum heap size 578, 586
changing 578
document processor 573
instance 581
changing 581
increasing 581, 585
search instance 593
changing 593
search server 593
setting 578
maximum log file size 610
maximum, indexer 586

memory consumption 572, 586
memory size 581
metadata 23, 55, 397–399, 492, 495
  extraction 360
  file system 494, 510
  format 402
  information 358, 361
mi category 292–293, 311
migration
  IBM Content Analyzer dictionary 308
  IBM Content Analyzer rule 313
modification of task route 511
monitoring
  parser status 598
  search server status 598
  system 595

N
Named Entity Recognition annotator 452
native ACL 637
native field 31, 33, 37, 39, 286, 398, 402
  mapping to search field 89, 103, 137
  updating, search field mapping 137
native source security 637
natural language processing (NLP) 3, 358
NLP (natural language processing) 3, 358
Node Id 574–575, 600
noun identification 299

O
OLAP Cubes 525
optional facet index 94, 591
OR operator 171

P
package.adic.xml 289
parse and index component 103, 106, 113, 122, 132
  starting 132
  stopping 133
parsed content 399
parser 598
  running with crawler and indexer 495
  status, monitoring 598
parsing content 493
Part of Speech (POS) 149, 225, 291, 310–311, 339, 341
  facet 54
password, expired 611
Pattern Matcher annotator 19, 35, 341–342, 452
  configuration 309
  disabling 310
  when to use 310
pattern matching 57, 291
  associating a Troubles facet 292
  association with a facet 292
  rule 292
  rules 338–339, 341
pattern, identifying 155
pattern-list, top-level element 311
PEAR (processing engine archive) file 141, 453
  single versus multiple 453
perfmon utility 602
performance
  facet querying 593
  tuning 596
performing an analysis 298
personal computer 4, 62
  help center 61
    staff 64
  overall customer satisfaction 64
  support ranking 64
Phrase Constituent 58, 149, 310, 341
plug-in
  architecture 561
  capability 446
  customize 562
  Java export 22
  template HTML file, change 563
  test 568
port assignment 75
POS (Part of Speech) 149, 225, 291, 310–311, 339, 341
  facet 54
precision 51
Predicate facet 334
predicate problems 330
Preferences link 147
Preferences window 150, 153
preservation of business data 507
problem 606
  detection 335
  identifying candidate expressions 331
  investigation 480
  predicates 330
  statement 2

process model 5
processing engine archive (PEAR) file 141, 453
  single versus multiple 453
processor utilization 572, 602
Product facet 156, 166, 226, 232, 284–285
product failure 4, 49, 57, 61
  early identification 53
  early warning 57, 61–63
property name 585
proposed clusters 348
proximity search 156
pruning the query tree 163
public security token replacement methods 637

Q
quality assurance 29
query
  condition 145, 154, 161–162
  facet path 156
  syntax 156–157
  tree 148
query builder 7, 180, 182
  best practice 193
  examples 186
query tree 7, 154–155, 163, 165
  access to 164
  best practice 193
  default query node 173
  editing 177
  examples 166
  left node 167, 170
  pruning 163
  right side 167
  understanding 165
  view 164

R
range facet 122
ranges within a range 121
ranking, search 364
Real-time NLP API 476
  when to use 479
rebuilding an index 137, 294, 573
recall 51
Records Management framework 522
Redbooks Web site
  Contact us xix
Regular Expression annotator 469
regulatory action 361
reindexing content 382
relational database 34, 400
  binary content 403
  crawler 17
  exported data 415
relevancy score 360
reliability correction 249
rendering 253
repair record 29
report generation 419
repository 485, 490–492, 637
  associated metadata 492
  fields 507
  IBM Content Analytics 492
  locations 492
resource deployment 135
resource preparation for Content Collector 508
restore command 140
result cache 590
role 635
role-based access 647
rule
  adding 311
  aggregate rules 457
  annotation 457
  creation of 311
  migration 313
  pattern file 294
rule-based categories 194
  configuration 195
  rule configuration 194

S
sales agent 66
sample data
  IBM Content Analytics 144
  Sample Text Analytics Collection 280
  small set 40
Sample Text Analytics Collection 144, 280–281
saved search 162
scheduling
  crawler 127
  deep inspection 441
  export 406, 424
search 358
  collection 11
  features 154

operator 154, 156, 163, 224
application 174
example 166
robust set 154
performing 422
ranking 364
runtime 11, 16, 20, 22, 24, 39–40
within a category 359
search box 145, 148, 151
current query condition 203
search expression 38, 145, 148, 153, 231
keyword portion 38
search field 32, 34, 38, 88, 103, 106, 148, 151, 281–282, 399, 494, 497
Clear button 203
creation of 96
large number 39
mapping to facet 90, 106
mapping to native field 89, 103
resulting list 37
updating, native field mapping 137
search result 23, 32, 38, 174, 203–204, 222, 309, 364, 395–396, 398, 406, 422
classifying 359
document 398–400, 408
document link 559
export 423
configuration 421
option 422, 428
requests 407
search server 41–42, 144, 572–573, 589–590, 595, 598
maximum heap size 593
status, monitoring 598
text analytics collection 42
searchable object 42
searching
and exporting a search result 423
Boolean operators 154
within a category 359
SearchService object 41–42
security 639
control, implementation 647
native source 637
security-token assignment 636
selected facet 55, 145, 218, 244, 303, 434
sudden increases 231
selected keyword 154, 245
actual frequency 242
semantic analysis algorithm 358
sensitive information 30
Service and Support Professionals Association (SSPA) 4
session
configuration 575, 579, 581
configuration ID 575
failed start 614
session ID 597, 600
session_ID variable 575
SIAPI 11, 20, 41–42
object-factory method 41
Social Security Number (SSN) 490, 497
sort criteria 235
source document 31
SPAN tag 567
sport people 47
SSN (Social Security Number) 490, 497
SSPA (Service and Support Professionals Association) 4
stand-alone validation tool 476
startccl command 579–580
statistical interval estimation 249
status
export request 407
of the crawler 131
stopccl command 578, 580–581
stopping a crawler 130
storage size, temporary 619
structured information 9
submitting a request of deep inspection report 438
survey 46
form 46–48
synonym 302–303
keyword 302
system log 608–609
changing log level 628
system monitor 595
system monitoring
AIX 602
Linux 602
Windows 602
system resource state 596
system scaling 73

T

TAE (Text Analysis Engine) 453
task route 508–510, 522–523

FSC Associate Metadata task 512
FSC Post Processing task 515, 519
IBM FileNet P8, Create the Document task 513
task route modification 511
tasklist command 602
taxonomy
browsing 359
cache 584
changing the default cache type 585, 587
Taxonomy Proposer 378, 385, 524
technical strength 68
technical trend 68
temporary disk space 610
temporary storage
increasing for crawled data 619
size 619
terms of interest 321, 326, 328, 330–331, 333–335, 338, 341
test plug-in 568
testing the UIMA pipeline 476
text analysis 4, 16, 20, 35, 46–47, 221, 299, 305, 309–310, 385, 486, 508, 523
critical element 299
Text Analysis Engine (TAE) 453
text analytics 1, 5–6, 33, 36, 48–50, 53–54, 490, 493, 497
collection 11, 21, 27, 31, 33, 90, 638
redeploying resources 475
resources, redeploying 475
Text Analytics tutorial 84
text miner application 16, 32, 294, 431, 523, 624
accessing 144
extend 560
log 608
logs 617
text miner application views 56
Deviations view 238
Documents view 219, 290
Facet Pairs view 244
Facets view 153, 222, 308
synonyms 303
Grid view 246
Time Series view 149, 153, 155, 226, 230
Trends view 155, 222, 231, 298
text mining 3, 16, 20, 56, 68, 152, 299, 311, 320, 358, 363, 384, 638
techniques 363
textual data 1–3, 10, 35, 45–47, 49, 51, 56, 59, 67, 69, 143, 153, 279–280, 387, 395, 398, 507
analytics 2
manual analysis 46
textual field 361
time period 162, 218, 226
time scale 226–227, 504
Time Series graph 231
Time Series view 149, 153, 155, 226, 230
time zone 425, 442
toolbar 145
application 145, 148
FSC Collector tasks 515
top-level element
dictionary 308
pattern-list 311
total number of documents
in a collection 170
in a corpus 203
that match a query 220
trace 605
trend 432
discovery 432
identification 68
identifying 155
Trends view 34, 55–56, 62, 155, 222, 231, 236, 298
trouble.pat file 292
troubleshooting 596, 619
general guidelines 605
tip 610
tuning the decision plan and knowledge base 383
TYPE_IO_ERROR message 619
type-ahead feature 157
configuration 157
Facet Navigation pane 159, 161
search box 158

U
UIMA
annotators 32
compliant 18
compliant annotator 458
document processing pipeline 450
PEAR 458
pipeline 18–19, 37, 382, 486, 579
annotator 19
IBM Classification Module annotator 382
testing 476
text analysis annotators 23
text analytics 37

Uniform Resource Identifier (URI) 559
unstructured content 1, 9
unstructured data 18, 32, 398, 486, 579
unstructured information 18–19, 23
URI (Uniform Resource Identifier) 559
use-case scenarios 28
  call center 28
  insurance fraud 28
  quality assurance 29
user data, location 574
user-defined Java routine 637

V
validation
  deep inspection request 444
  export request 426, 430
  techniques for custom annotators 482
valuable insight 3
verification of installation 82
view
  log content 609
  query tree 164
view tab, change title 563
vmstat commands 602
VOC (voice of customer) 2, 60
voice of customer (VOC) 2, 60

W
warning of product failure 63
web application server
  Jetty 647
  WebSphere Application Server 647
web crawler 578, 597, 637
  real-time monitor 578
web-based crawler 16
WebSphere Application Server 647
WebSphere portal crawler 17
wildcard search 156
window capture 252, 606
Windows
  file system crawler 123
  system monitoring 602
word cloud 562
word, intent of 358
workflow of content decommission 491

X
XML
  attributes, exporting 405
  data 280
  element 281, 286, 402, 405, 410, 510
    name 281
  file 24, 363, 367, 402, 432, 502, 506, 513, 559
    format 432
    import to Classification Module 355
    wizard 510
  format 22, 402, 433, 445, 597
    output files 445
  version 280, 292, 413–414, 426

IBM

Redbooks

IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

IBM Content Analytics Version 2.2

Discovering Actionable Insight from Your Content

Learn about Content Analytics and its value to your organization

See how to discover actionable insight from your content

Understand key product features and offerings

With IBM Content Analytics Version 2.2, you can unlock the value of unstructured content and gain new business insight. IBM Content Analytics Version 2.2 provides a robust interface for exploratory analytics of unstructured content. It empowers a new class of analytical applications that use this content. Through content analysis, IBM Content Analytics provides enterprises with tools to better identify new revenue opportunities, improve customer satisfaction, and provide early problem detection.

To help you achieve the most from your unstructured content, this IBM Redbooks publication provides in-depth information about Content Analytics. This book examines the power and capabilities of Content Analytics, explores how it works, and explains how to design, prepare, install, configure, and use it to discover actionable business insights.

This book explains how to use the automatic text classification capability, from the IBM Classification Module, with Content Analytics. It explains how to use the LanguageWare Resource Workbench to create custom annotators. It also explains how to work with the IBM Content Assessment offering to timely decommission obsolete and unnecessary content while preserving and using content that has business value.

The target audience of this book is decision makers, business users, and IT architects and specialists who want to understand and use their enterprise content to improve and enhance their business operations. It is also intended as a technical guide for use with the online information center to configure and perform content analysis with Content Analytics.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7877-01 ISBN 0738435287